Mystery on the Bilingual Express: A Critique of the Thomas and Collier Study "School Effectiveness for Language Minority Students"

Christine H. Rossell

Perhaps no other "yet to be released" report has been quoted so much or so often as the so-called "Collier Study."[1] Approximately two years before the report was completed, Virginia Collier was holding public meetings at which she disseminated a five page summary of her "study" (Thomas and Collier, 1995)—two pages of text, two pages of line graphs, and a one page list of program definitions. In no time, the "Collier Study" had become another factoid in the controversy over bilingual education. Even though no one had actually read it, the report was being cited everywhere as proof that bilingual education, particularly two-way bilingual education, was superior to all other programs for limited-English proficient (LEP) children.

Some two years later, the complete report has finally been issued. It can be downloaded from the National Clearinghouse for Bilingual Education web page at http://www.ncbe.edu. Although 96 pages long, it contains no more data on the findings of the study than the same two charts in the original press release. There are no tables at all and the few other charts in the study are merely illustrations of their theories. In fact, this report consists primarily of theories of bilingual education and criticism of the scientific method.

The methodology of the study is unscientific, as is the case with all of Virginia Collier's research. The criteria for a scientific study (see Rossell and Baker, 1996a, 1996b) are basically four-fold. First, there should be a treatment group—for example, LEP students in a bilingual program—and one or more comparison groups—for example, similar LEP students in one or more types of all-English programs. Second, the achievement of these students should be compared after some time period in their respective programs. Third, any differences between the students initially should be controlled for statistically in order to give each group a level playing field. (This is not necessary if there is random assignment.) Fourth, the *same* students must be followed over time since there is no way to statistically control or match on initial differences, nor would it make any sense to do so if different students are in the study at different points in time. Although all four characteristics are essential, only the first two are found in the Thomas and Collier study.

A treatment and a comparison group are necessary in order to interpret outcomes. If students in a bilingual education program score at the 30th percentile, it is a positive effect if they would have scored at the 20th percentile in an alternative program, a negative effect if they would have scored at the 40th percentile in an alternative program, or no effect at all if they would have scored at the same 30th percentile in an alternative program. It is only this comparison of students in a bilingual program to students in an alternative program that enables us to evaluate what a score at the 30thpercentile means.[2]

But comparing students in alternate programs is not enough. One must also statistically control for any pre-treatment differences between the two groups. If students are randomly assigned to different programs, a statistical control for pre-treatment differences is not necessary because we can be sure that any difference between the outcomes of the two programs is not a result of the characteristics of the students.

However, random assignment is rarely possible in the real world. In the real world, students are assigned to, or select themselves for, different programs based on their individual characteristics such as motivation, intelligence, social class, or learning problems. These differences that exist before the program will be confused with the effects of the program unless statistically eliminated. For example, if the students in a bilingual program are poorer than the students in an alternative program, it would be unfair to compare the two groups without adjusting statistically for these differences in social class.

It is also essential that the same students be followed over time because otherwise we have no way of knowing whether the students were initially comparable before the program. Nor would it make any sense to control for pre-treatment differences if there are different students before the treatment than there are after the treatment.

Virginia Collier's and Thomas and Collier's studies have some, but not all, of these essential characteristics. But rather than apologizing for this and cautioning the reader, as is the professional norm in our field, they attempt to discredit the scientific method. Some of their criticisms are valid and some are not. As they point out, the typical scientific study examines growth in achievement over a short time period because of the requirement to follow the same students over time. But where Thomas and Collier go wrong is that they propose to "solve" this problem by following *different* students over eleven grades, and to have no statistical control for pre-treatment differences. This is analogous to criticizing the judicial system for its failure to consistently result in conviction for the guilty and acquittal for the innocent and offering as the solution throwing the accused in a vat of water—if they are innocent they will float, otherwise they will sink.

Their other criticisms are also problematic. For example, they criticize Rossell and Baker (1996) for suggesting that "only studies in which students are randomly assigned to treatment and control groups are methodologically acceptable" (p. 20). In fact, that was never said. Of the seventy-two methodologically acceptable studies that we identified, only five had random assignment. The other sixty-seven were quasi-experimental—that is they had a treatment and a control group, as well as a control for pre-treatment differences in the absence of random assignment. Since a quasi-experimental design is both scientific and "do-able," Thomas and Collier really have no excuse for not using it in their study.

They also erroneously assert that the scientific method does not work because "At best, one might find a comparison group that received an alternative form of special assistance, but even this alternative is not easily carried out in practice" (p. 20). In fact, it is simple to find a comparison group because the vast majority of LEP children in the U.S. are in alternative programs. The California Department of Education language census, for example, (http://www.cde.ca.gov—Table 13) indicates that in 1997 only 29.7 percent of the LEP students in the state were enrolled in bilingual education. Even in Los Angeles, only 34 percent of the LEP students are enrolled in bilingual education. Thus, Thomas and Collier are wrong—it is easy to find a comparison group that has received an alternative form of special assistance and it is often possible, "to carry this out in practice" if one is trained in social science research methods.

There are more errors in their criticism of the scientific method. On page twenty-two, they state that "If truly comparable local control groups are not available, one can construct a comparison group from the performance of other groups such as the norm group of a nationally normed test." Unfortunately, one *cannot* construct a comparison group from the norm group of a nationally normed test. First, LEP children are defined as LEP precisely *because* they score below the national norming group of fluent English speakers, and so the national norm group will always be higher. For example, a common criterion for classification as LEP is scoring below the 36th percentile and a common criterion for reclassification as no longer LEP (i.e., Fully English Proficient [FEP]) is scoring above the 36th percentile. Thus, no matter how effective a program is, children classified as LEP will by definition score below the average for fluent English speakers. Second, the *growth* in achievement of LEP children is much greater than the growth in achievement of FEP and native speakers of English because the tests are normed on a fluent English speaking population. If fluent English speaking children are making grade level progress they will leave a grade with the same score with which they entered. LEP children, by contrast, may make greater progress than this because they are getting two treatments: learning the English language and learning the subject on the test. Thus, the growth in the achievement of the two groups is not in any way comparable. Third, only half of all LEP children will reach the national norm because only half of all English monolingual children ever reach it. How is one to tell if the national norm is the right standard for any individual child? For some children it may be too high and for others too low. In short, a national norm group is not the right criterion by which to evaluate bilingual education. That Thomas and Collier continue to perpetuate this myth is both a major problem with their study specifically and with the field of bilingual education generally.

WHAT IS THE TREATMENT?

Thomas and Collier criticize current research because programs vary greatly in how they are implemented and it is difficult to tell exactly what is being evaluated in these studies. This is true, and their study has the very same problems because it is massive and covers a fifteen year period. In their own words, the study

> includes over 700,000 language minority student records, collected by the five participating school systems between 1982 and 1996, including 42,317 students who have attended our participating schools for four years or more. This number also includes students who began school in the mid-1970s and were first tested in 1982. (p. 30)

There is no way they can know exactly what is going on in these school districts because rather than doing their own research, they relied on school staff to define the programs. Although this is what most researchers do, it is a mistake. Having visited almost a hundred classes in my research on bilingual education, I have learned that not only can you *not* rely on school staff to define programs accurately, you cannot even rely on the teacher in the classroom to define her own approach accurately (see Rossell and Baker, 1996b, Chapter 4). Even today teachers will claim to be teaching bilingually in programs *currently* labeled bilingual education when actually they are teaching completely or almost completely in English. You absolutely cannot expect teachers to reliably or accurately describe programs implemented fifteen years ago, even if you could find someone who had been in the school system that long.

There are, of course, problems with all studies, but what is striking about Thomas and Collier is their failure to acknowledge that their study has these or any other problems. Indeed, quite the contrary, they believe their own study overcomes the limitations of previous research. They are wrong.

PURPOSIVE SAMPLING

Thomas and Collier acknowledge one limitation when they admit that

> our findings are generalizable only to well-implemented, stable programs from school systems similar to those in our study. This is not accidental. We intended to select a purposive sample of above-average school systems...all of which were well implemented by experienced, well-trained school staff. (p. 28)

However, they never define "well-implemented," "stable," "above-average," "experienced," or "well-trained." Indeed, no data are presented on any of the characteristics of any of these programs, the children enrolled in them, the schools in which they reside, or the school districts. *We literally know nothing about these school districts and their schools* other than the fact that there are five of them and they are "moderate-to-large, urban and suburban school systems from all over the U.S." (p. 30). Even their programs are defined only in broad generalities that could apply to any program of that type in any district. Since the schools and school districts are not named, there is no way to check their assertions that these are well-implemented programs, or even that they are *equally* well-implemented programs—that is, well-implemented across program types.

Even if it were possible to define "well-implemented," how could one do so across eleven grades and a fifteen-year time period from 1982 to 1996? Thomas and Collier never tell us. They collected information on program quality and classroom atmosphere by holding "focus groups." They explain this process as follows:

> To measure school program effectiveness in each school district, we began by interviewing school staff to identify and reach a consensus on definitions of programs and their implementation. We did this through focus groups with bilingual/ESL teachers and resource staff. Through these group interviews, we uncovered differences from one school dis-

> trict to another in the labels given to programs, but consistency in general characteristics of differences between programs. We have chosen here to report these findings by using the names of general program labels in bilingual/ESL education. (p. 48)

We do not know what questions were asked or what topics were discussed. Nor do we know when during the fifteen-year period these focus groups were held, how many were held, and how information obtained in this process helped to define programs and evaluate the "atmosphere." In the determination of how "bilingual" a program is, focus groups are neither a proper nor an adequate substitute for observations of classrooms. They are also not a substitute for hard data on program quality. These data might include class size, teacher qualifications, percentage of English used in instruction, racial composition of the classroom and school, percentage eligible for free or reduced lunch in the classroom and school, percentage classified LEP, and so forth. None of this is in the report, and there is no promise of it in future reports. In fact, in twenty-five years of reading technical reports, I do not think I have ever read one with so little information in it. Certainly, I have never read a federal grant report with so little information.

The last federal grant report I wrote (Steel, Levine, Rossell, and Armor, 1993), had seventy-four tables, fifty-nine charts, and 134 pages of appendices explaining the sampling procedure, the characteristics of the sample, and the methodology. Approximately 37 percent of the report was devoted to tables and charts, and 68 percent of the report was devoted to tables, charts, and appendices explaining sampling procedures, sample characteristics, and methodology. Our appendices also included the complete text of the questionnaire that we used. Another federal grant report, this one by Ramirez et al. (1991), contained even more tables and figures than ours, although proportionately it was about the same. This two-volume report contained 1,148 pages of which there were 478 tables and 228 figures on the characteristics of the sample, the outcomes, and the statistical analysis. The methodology was explained in intricate detail, and the tables and figures represent about sixty-one percent of the report.

The Thomas and Collier report by contrast has no tables at all, no information on the characteristics of the sample and methodology, and no statistical analyses. The two figures represent only one percent of the report. This appears to be a new low in federal grant reporting.

METHODOLOGY

Because of the lack of information in the report, it is hard to know exactly what Thomas and Collier have done. As a result, this has generated some

confusion among experts in bilingual education. One prominent expert who has read the study believes their line graphs reprersent not real achievement, but *predictions* of what these students would achieve. This assumption is fueled by the fact that the test scores cover eleven grades, but the longest study that anyone knows about is five years—(the Ramirez et al. [1991] study which is apparently no longer in their database).[3] What people don't realize is that this is not a true longitudinal study because most of the students in the sample have only four years of data, and there are almost no students with eleven grades of data. Thus, each grade consists of different students. Rather than apologize for this, they brag about it. Indeed, they claim to

> have greatly increased the statistical power of our study with very large sample sizes. We have achieved these sample sizes, even when attrition reduces the number of students we can follow over several years, by analyzing multiple cohorts of students for a given length of time (e.g., seven years) between major testings. (p. 28)

Thomas and Collier do not tell us the size of each cohort, their initial test scores, their ethnicity, their social class, nor how they were matched—that is, what leeway was allowed in the matching and at what point in their education. The explanation for their procedure is provided in only a few paragraphs:

> The sample figure below illustrates eight available seven-year testing cohorts for students who entered school in Grade 1, were tested in Grade 4, and who remained in school to be tested in Grade 11.
>
> We then analyzed multiple cohorts of different students over a shorter time period (e.g., six years), followed by successive analyses of different students in multi-year cohorts down to the four-year testing interval. In doing this, we have in effect "modeled" the typical school system, where many students present on a given day have received instruction for periods of time between one and twelve years. Typically, the shorter-term cohorts (e.g., four years) contain more students than the longer-term cohorts since students have additional opportunities to leave the school system with each passing year.
>
> Using this approach, we are able to "overlay" the long-term cohorts with the shorter-term cohorts and examine any changes in the achievement trends that result. If there are no significant changes in the trends, we can then continue this process with shorter-term cohorts at each stage. If significant changes occur in the data trends at a given stage, we pause and explore the data for possible factors that caused the changes. (pp. 28–29)

The above paragraphs are pretty much the sum total of the explanation of their methodology (although they repeat it again on pp. 53–54), and they are

Table 1
Years and Grades of Test Administration

	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
Cohort 8								4		6		8			11
Cohort 7							4		6		8			11	
Cohort 6						4		6		8			11		
Cohort 5					4		6		8			11			
Cohort 4				4		6		8			11				
Cohort 3			4		6		8			11					
Cohort 2		4		6		8			11						
Cohort 1	4		6		8			11							

Source: Thomas and Collier (1997: 29).

simply inadequate. Moreover, the first sentence—that students entered in first grade were tested in fourth grade and remained in school until eleventh grade—contradicts the fourth sentence that says there were shorter-term cohorts of four years and that these contain more students than the longer-term cohorts. In addition, Table 1 shows exactly eight years of data for each cohort. Where are the four-year cohorts that Thomas and Collier refer to? Where are the five-year cohorts, and so forth? Why do Thomas and Collier refer to the cohorts as seven-year cohorts when eight years of achievement data are portrayed in the table?

It appears that the cohorts were defined by the year (1982, etc.) they started analyzing achievement data. Each cohort, however, is comprised of different students tested at different times—that is, they "overlay" the long-term cohorts with the shorter-term cohorts. Thomas and Collier never tell us how many students were followed over four years and what percentage of the total program enrollment they represent, nor how many students were followed over five years and what percentage of the total program enrollment they represent, and so forth. There is just a vague reference to the fact that there are more students in the shorter-term cohorts.

Their discussion of matching students is also inadequate. They claim to be

> "blocking," first to group students using categorical or continuous variables that are potential covariates, and then later to use these groups as another independent variable in the analysis. Essentially, all student scores that fall into the same group are considered to be matched (Tabachnick & Fidell, 1989, p. 348) and the performance of each matched group within each level of program type can be compared. Each group can then be followed separately and its performance on the outcome variables (typically test scores) can be investigated separately from that of other groups of similar students. Interactions between the new independent variable represented by the blocked groups and other independent variables (e.g., type of program) can be investigated.... However, when the assumptions of ANCOVA [analysis of covariance]could be met, we used ANCOVA as a supplement to blocking, in order to take advantage of the benefits of both techniques in situations where each works best. In some of our analyses, we used an expanded, generalized form of ANCOVA. (p. 27)

But what scores did they match on? Which groups of students were matched and how? Their charts do not follow each group separately and they never give us this information. Nor is there a single statistical analysis in this report. When and where did they use ANCOVA, and what happened to the results? Why did they not present them?

Number of Years in the Program

Thomas and Collier ignore the number of years students spent in the program. They justify this on the grounds that ESL pullout programs are designed to be short-term, limited instructional support: "There are no four year, five year, or six year ESL-pullout programs in existence in our participating school systems" (p.54). However, in every one of the school systems I have personally analyzed, students were in ESL pullout programs until they were re-designated. Since the criterion for re-designation was often a percentile criterion that was not achievable by a third or a half of all English monolingual students (e.g., the thirty-sixth or fiftieth percentile)—let alone students who are still learning English—some students received ESL support their entire elementary school careers. This is also true of so-called early exit bilingual programs—that is, some students exit after two or three years and some are in them for their entire elementary school careers. Thus, not controlling for length of time in the program introduces error into an analysis.

Validation of Results

Thomas and Collier assure the reader that they "validated" the findings from their five participating school systems by "visiting" other school systems in twenty-six U.S. states during the past two years and asking the school districts to "verify" their findings. But they never tell us what "validate" and "verify" mean. Does it mean having conversations with people? Does it mean checking the statistics of other school districts to see if they are similar? If so, how similar would they have to be to qualify as "validation"? Since school districts typically do not keep longitudinal data like this (and if they do, it is not readily accessible), it is hard to believe it is the latter. We will never know what "validate" and "verify" mean because Thomas and Collier just do not think it is important to tell us.

DISCARDING UNWANTED RESULTS

Although Thomas and Collier state that "each school district's data [is] analyzed separately" (p. 30), they do not explain what that means. The chart shows six lines for five school districts. How many programs per school district are there? If the five school districts are analyzed separately, why are there only six lines?

They apparently discarded results they thought were not "generalizable":

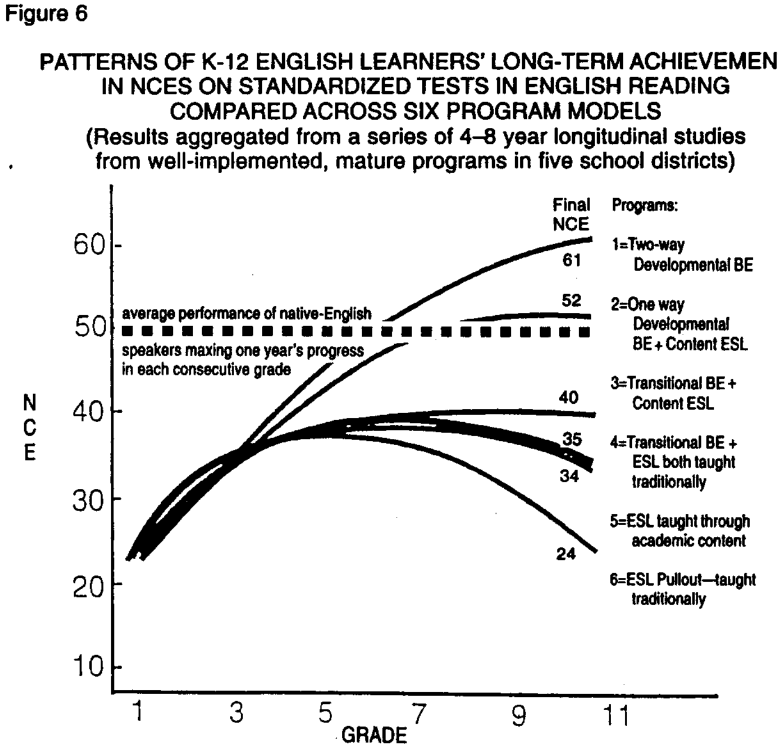
> We first examined patterns in one school district, to see if any program or instructional variables appeared to have strong influence on language minority students' achievement. Then if a *particular pattern* emerged, we assumed it was not generalizable beyond the context of that school system, unless we found a *similar pattern* in a second school system. Once the *same pattern* appeared *repeatedly* in the data across more than two research sites, we *started to assume some generalizability*. The patterns that we are repeating here are general academic achievement patterns across all five of our research sites. [Emphasis added] (p. 48)

Discarding results you don't like is an unacceptable approach to data analysis. It also means that this study can never be replicated since everyone will have a different perception of such words and phrases as "similar pattern," "repeatedly," and "started to assume some generalizability."

NEW DATA, SAME TRENDS

Another oddity is that the trend lines in Figure 1 in this report appear to be the same as those in Thomas and Collier's 1995 press release (despite the fact that the Ramirez et al. (1991) study was apparently dropped from their database and two more years of data added). How does this happen? Well, if the researchers think it is perfectly fine to pick and choose among achievement trends, selecting only those that "assumed some generalizability," how could it not happen? Any new data that deviates from the earlier pattern is discarded because it is not "generalizable." Thomas and Collier have thus taken a giant step backwards from the scientific goal of establishing research standards that are objective and verifiable. Once again, they do not even apologize for it. Quite the contrary, they think this is progress!

Figure 1
Thomas and Collier's Chart of Elementary School Programs



Source: From *School Effectiveness for Language Minority Students* (p. 53) by W. P. Thomas & V. Collier, December 1997, NCBE Resource Collection Series, No. 9. Washington, DC: National Clearing House for Bilingual Education. Copyright by Wayne P. Thomas and Virginia P. Collier, 1997. Reprinted with permission from NCBE and Thomas and Collier. http://www.ncbe.gwu.edu/ncbepubs/resource/effectiveness/thomas-collier97.pdf

OVERCOMING POVERTY

Thomas and Collier claim that "the schools with the highest achievement levels were so effective that the effect of these programs overcame the power of student background variables such as poverty" (p. 48). But the only analysis that would justify such an assertion would be one showing that children of lower social class were scoring the same or higher than children of higher social class controlling for the students' years in this country, years in their programs, age, and initial language fluency. Since they apparently did no statistical analysis of this kind, they cannot know that the programs have overcome the effects of poverty. If these programs had actually overcome the effects of poverty they would be the first in the four decades of social experimentation in the U.S. and the world would be beating a path to their door.

First Tested in Fourth Grade

According to Thomas and Collier, the students in their study were first tested in fourth grade. This may be true of the students in bilingual education, but it would be an oddity to have a student in an ESL program not tested until fourth grade. It is equally odd, however, that although the students were apparently not tested until fourth grade, Thomas and Collier's elementary graph in Figure 1 shows test scores that average approximately 22 in first grade and 35 by third grade. Where do these data come from? How can students who enter school with no proficiency in English be scoring at the twenty-second percentile by first grade regardless of the type of program they are in? Why did the 1995 chart have the same average score of 22 in *kindergarten* when presumably no students had been tested? Thomas and Collier do not explain any of these issues.

Only Reading Achievement?

Another of the many mysteries in this report is why the authors only examined reading achievement. They claim this is justified because reading involves more complex problem-solving across the curriculum than do language arts. However, they are silent on why they did not also analyze math test scores, presumably a subject involving complex problem-solving.

It is interesting that the two achievement areas that Thomas and Collier failed to analyze in their report—language arts and math—are the two areas where bilingual education does the worst. In our review of the research on bilingual education (Rossell and Baker, 1996a, 1996b), Keith Baker and I found that transitional bilingual education was *worse* than "doing nothing" (also called submersion) only 33 percent of the time in reading achievement, but 64 percent of the time in language achievement, and 35 percent of the time in math achievement. Transitional bilingual education was *better* than doing nothing 22 percent of the time in reading achievement but only 7 percent of the time in language achievement and nine percent of the time in math achievement. In short, although Thomas and Collier try to justify this on other grounds, they analyzed only the subject area where bilingual education is most successful.

Attrition Bias

Thomas and Collier claim to be studying the best implemented programs. Regardless of whether this is true, they *are* studying the best students since these are the only ones with test scores over a four-year period. This is called *attrition bias*, and it is a problem that all longitudinal studies have. It can lead to misleading conclusions if the program is only effective for the most stable people but not for the whole population in the program. As with many issues in this report, Thomas and Collier are silent on the extent to which attrition bias might be a problem in their study.

Non-testing of Bilingual Education Students

There is a huge differential testing bias in bilingual program evaluations that I have only recently become aware of—a much smaller percentage of the bilingual education students are tested than of the all-English LEP students. For example, the Los Angeles Unified School District recently reported higher achievement for students who had stayed in the bilingual program for five years compared to similar students in the mainstream, but almost 40 percent of the students in the bilingual program were thought to not know enough English after five years to be able to take the test. By contrast, 97 percent of the all-English program LEP students took the test (Zacaria, 1998).

Recent data from a fifty-four percent Hispanic school district of 17,000 students in California shows a similar pattern. In this district only 36 percent of the Spanish speaking LEP students, almost all of whom were in bilingual education, were tested, but 83 percent of the non-Spanish speaking LEP students, all of whom were in all-English programs, were tested. Similarly, in the Ramirez et al. (1991) study, 89 percent of the immersion strategy students were tested in the K–1 analysis, but only 61 percent of the early exit students were tested. In the grades 1–3 analyses, 42 percent of the immersion strategy students were tested, but only 29 percent of the early exit students were tested.[4] That study found no difference between the two programs, but this obviously underestimates the benefit of immersion and overestimates the benefit of bilingual education since it is the poorest students who are not tested and there are more untested students in the bilingual education program.

Enrollment in Each Elementary School Program

To reiterate, from over 700,000 student records, Thomas and Collier were able to identify 42,317 student records in four-year, five-year, six-year, and so on up to eight-year *overlapping* testing cohorts to demonstrate a "longitudinal perspective." Thus, the records analyzed actually represent 6 percent of the total database and lots of different students at each point. Each line has an underlying long-term longitudinal cohort, with a series of overlapping shorter-term longitudinal cohorts. Each line in the graph represents a weighted average (which they do not explain) of all the cohort scores available at each grade level. The only data presented in the report are the six lines representing six programs in Figure 1 (Figure 6 in their report) and the two lines in Figure 8 in their report, for two high school ESL programs. (There are no students in bilingual education during high school.)

Although Thomas and Collier do not give us information on numbers of students in each cohort or program or grade, they do provide us with overall percentages. Three percent of the students were in two-way bilingual education (line one in Figure 1); 7 percent were in one-way developmental (or maintenance) bilingual programs (line two); 9 percent were in transitional bilingual education with ESL taught through academic content (line three); 17 percent were in transitional bilingual education taught "traditionally," which apparently includes ESL taught "traditionally" (line four); 13 percent were in ESL content (line five); and 51 percent were in ESL pullout (line six). These percentages also refute Thomas and Collier's claim that it is difficult to assemble a comparison group since 64 percent of the students in their sample were in a comparison group.

These overall percentages raise as many questions as they answer because Thomas and Collier do not tell us how they were calculated. The calculation of percentages is not trivial in a study that consists of different students in each grade. Furthermore, with only 3–7 percent of the students enrolled in the developmental bilingual education programs, one can only wonder how many students actually had eleven years of data—it is not inconceivable that there were virtually none. This would explain why no sample sizes are shown for each grade and program.

Program Definitions

Two-way Developmental Bilingual Education

Thomas and Collier define two-way bilingual education as follows: "language majority and language minority students are schooled together in the same bilingual class, and they work together at all times, serving as peer teachers" (p. 58). Both the 90–10 (ninety to one hundred percent Spanish in Kindergarten and first grade, reaching fifty/fifty by fourth or fifth grade) and the 50–50 (Kindergarten to fifth grade, half a day in Spanish) are included in the two-way bilingual education models. Thus, programs using very different amounts of Spanish are lumped together.

Many people who have looked at these data, including advocates of bilingual education, believe that Thomas and Collier combined the test scores of native speakers of English and limited English proficient students. Most people are willing to say it only verbally, but Kathryn Lindholm has put it in writing. She cites Collier's discussion of bilingual education theory (1995) and the findings of the Thomas and Collier study of 1995 for the proposition that

> ...the results of the Spanish and English speakers are aggregated rather than separated by language group (Christian, Montone, Lindholm & Carranza, 1997; Collier, 1995). (Lindholm, 1998:4)

Yet, Thomas and Collier do not admit to this, and their brief discussion states that they only looked at limited English proficient students:

> ...It is important to remember that this figure represents cohorts of students who start school with the same general background characteristics—i.e., no proficiency in English and low socio-economic status as measured by eligibility for free or reduced lunch. (pp. 53–54)

School districts that do break data down by Spanish and English native speakers have the following test scores for Hispanic and white students in two-way immersion programs.

Table 2 contains data from the Hernandez two-way bilingual programs in Boston. Where there is no data means there were less than seven students who took the test (or knew enough English to take the test). In grades one through three, apparently less than one-third of Hispanics in the school knew

Table 2
Reading Test Scores (Metropolitan) for Hispanic and White Students at Hernandez Two-way Bilingual School, Boston, MA, May 1993

		GRADES							
		1	2	3	4	5	6	7	8
Hispanics:	Reading Score	—	—	—	43	30	30	25	39
	Enrollment	*(26)*	*(26)*	*(25)*	*(25)*	*(22)*	*(19)*	*(10)*	*(15)*
Whites:	Reading Score	—	88	98	—	—	—	—	—
	Enrollment	*(5)*	*(9)*	*(7)*	*(5)*	*(6)*	*(2)*	*(0)*	*(3)*

Note: No data means less than seven students took the test.
Source: Boston Public Schools (1994).

Table 3
Reading Test Scores (CAT 5) for Hispanic and White Students by Language Proficiency at the River Glenn Two-way Bilingual School, San Jose, CA, Spring 1997

		Language Fluency Classification			
		LEP	FEP	English Only	TOTAL
Hispanics:	Reading Score	20	41	33	
	Enrollment	*(150)*			(283)
Whites:	Reading Score	—	—	65	
	Enrollment	*(0)*	*(0)*	*(113)*	*(113)*

Source: San Jose Unified School District (1998).

Table 4
Reading Test Scores in Grade Equivalents (CAT) for Hispanic and White Students in the Amigos Two-way Bilingual Program, Cambridge, MA, Spring 1991

		Grades		
		1	2	3
Hispanics:	Reading Score	1.29	3.11	2.87
	(% Tested)	*(46%)*	*(67%)*	*(47%)*
Whites:	Reading Score	1.30	5.09	4.70
	(% Tested)	*(67%)*	*(100%)*	*(86%)*

Source: Cazabon, Lambert, and Hall, (1991), p. 8, 14–16.

enough English to take the test. In the later grades it might be only slightly more than that. The Hispanic students who know enough English to take the test in this well regarded, two-way bilingual program are scoring at the 39th percentile even as late as eighth grade. But virtually all English speakers take the test and they do superbly—average scores between the 88th and 98th percentiles.

Table 3 shows the results for Hispanic and white students by language proficiency at the River Glen Two-way Bilingual School in San Jose, California. The Hispanic achievement in this two-way bilingual program is similar to the Hernandez School—it ranges from the 20th to the 41st percentile. Across all language proficiency levels, Hispanic students in this school are low scorers, about half that of the whites.

Table 4 shows English reading achievement data expressed in grade equivalent scores for Hispanic and white students in the Amigos Two-way Bilingual program in Cambridge, Massachusetts. In the spring of the year, a grade-level score (i.e. the 50th percentile)[5] for first grade would be 1.7, for second grade it would be 2.7, and for third grade it would be 3.7. The Hispanic students in this two-way bilingual program are below grade level at each grade. The whites are also below grade level in English in first grade when most of the instruction is still in Spanish. By second grade they are well above grade level and have twice the achievement of the Hispanic students.

As these data make clear, the Hispanic students in well-regarded, two-way bilingual programs in real school districts score only about half as well as white students and well below the Thomas and Collier Hispanic achievement data, even though only the top Hispanic students are tested. Even in the Hernandez School in Boston, which has data through eighth grade, Hispanic test scores are at the 39th percentile, not the 55th that is shown in Thomas and Collier's graph. Since their data do not match the Hispanic achievement found in other well-regarded, two-way bilingual programs, it is not surprising that people believe Thomas and Collier have combined the test scores of whites and Hispanics.

One-way Developmental Bilingual Education

Historically referred to as maintenance bilingual education or late-exit bilingual education, Thomas and Collier define one-way developmental bilingual education as "academic instruction half a day through each language for Grades K–5 or 6" (p.58). Originally planned for Kindergarten through twelfth grade, it has rarely been implemented beyond elementary school level in the U.S. Although one-way developmental bilingual education programs are nine points below the two-way programs, they are twelve points above transitional bilingual education. Some of this advantage for developmental programs (one-way and two-way) might be self-selection bias.

Thomas and Collier present no statistics on the social class or English fluency of the students enrolled in these programs. However, published reports from school districts with two-way programs indicate that developmental bilingual programs are likely to have significant numbers of fluent-English-speaking Hispanics enrolled in them in order to develop their Spanish. For example, according to a survey of students conducted in the Cambridge Amigos program, 41 percent of the fifth and sixth grade Spanish Amigos and 53 percent of the fourth grade Spanish Amigos hardly ever or never speak Spanish at home with their parents or brothers or sisters. This suggests that they started the program as fluent English-speakers and come from homes where English is the medium of communication.

In addition, the Hispanic students in developmental programs tend to be of higher social class. In San Jose, for example, only 29 percent of the Hispanic students in the River Glenn two-way bilingual program are eligible for free or reduced lunch compared to the district percentage for Hispanics of 67 percent.

Transitional Bilingual Education (TBE) Plus Content

Also called "early-exit bilingual education," Thomas and Collier define this approach as "academic instruction half a day through each language and gradual transition to all-majority language instruction in approximately 2–3 years" (p. 58). Again, I quarrel with their definition. The early exit programs I am familiar with teach almost completely in Spanish in kindergarten. In first grade they teach reading and writing in Spanish and all subject matter in Spanish with only an hour or so a day of English. In second grade, some subject matter may be taught in English but only if there are no new non-English-speaking students. In short, the first two years of these early-exit programs are almost completely in Spanish, although their goal is to transition the students to an all-English program as soon as possible. Thus, my experience is that the major difference between developmental and transitional (or early-exit) bilingual education in the first two years is the goal, not the amount of English.

English-as-a-Second Language (ESL)

Thomas and Collier describe ESL as coming in two forms:

1. ESL academic content, taught in a self-contained classroom (also referred to as sheltered instruction or structured immersion) which varies from a half-day to a whole-day; and
2. ESL pullout which varies from 30 minutes per day to a half day or at the secondary level from one to two periods per day.

In Figure 1 (Figure 6 from their report), line five is ESL academic content and line six is ESL pullout.

Why a Descriptive Cohort Analysis is Not a Methodologically Acceptable Design

None of the major research reviews have accepted a descriptive cohort analysis of the type Thomas and Collier offer as scientific evidence of the effectiveness of a program. A descriptive cohort analysis is generally unacceptable because it consists of different students at each point in time and there is no control for student characteristics.

In Appendix 2, I demonstrate how it is mathematically possible for the average achievement for a bilingual program to increase from fourth through eleventh grade (eight years) even though each individual student in the program experienced a decline over the four years of testing. I produced this phenomenon by doing just what Thomas and Collier did—I overlaid a series of about ten or so sets of four years of hypothetical achievement data beginning and ending at different points in the eleven year grade-span for each program.

I was able to do this so that each individual cohort had the opposite pattern of the overall average for each grade. For example, each hypothetical student in the two-way bilingual education program had a decline in achievement over their four or so years of data, but the overall average for each grade in the bilingual program went up dramatically. Conversely, each hypothetical student in the ESL pullout program had an increase in achievement over their four years or so of data, but the overall average for each grade in the ESL program went down across the whole eleven years. This mathematical phenomenon is well known among social scientists trained in statistics, and it is why the methodology used by Thomas and Collier—a simple descriptive cohort analysis—is not a valid means of evaluating programs and will continue to be rejected by scientists.[6]

Secondary School Achievement

Thomas and Collier do not analyze bilingual education programs at the secondary level because they feel there are not enough years to warrant analysis. As with the chart on elementary school programs, they present no data whatsoever on the school districts, schools, and students, and there are no statistical analyses of any kind. There are two programs analyzed—ESL through academic content (sheltered subjects) and ESL with traditional approaches, shown in Figure 8 of their report. Again, they examine only reading achievement. The descriptive cohort analysis shows these students apparently began school in fifth or sixth grade and by the eighth grade were scoring at the 20th percentile. By eleventh grade, they are at the 36th percentile in ESL through content and the 25th percentile in ESL taught traditionally. All of the important problems identified in the elementary school program analysis apply to the secondary school analysis.

CONCLUSIONS

The Thomas and Collier study has two serious flaws. First, it uses a methodology—a simple descriptive cohort analysis—that is unscientific and that can produce misleading results as I have demonstrated in Table 5. The method is unscientific because each grade consists of different students, the number of students in each grade is not given and could be quite small, there is no statistical control for pretreatment differences, the programs are elementary school programs, but the test scores shown after elementary school are of different students from those in the elementary programs. Even if they were the same students, it would be hard to imagine how a program that a student participated in years earlier would influence him or her without controlling for the student's characteristics and the characteristics of his or her current school and program. Equally as egregious is the fact that Thomas and Collier have admitted they selected only the trends that they thought "assumed some generalizability."

The second serious flaw in the Thomas and Collier report is that they explain very little about their methodology. They present no data whatsoever on their sample or any statistical analyses. Normally a federally funded grant would have a final report that included at a minimum, dozens of tables in the text and in appendices with the characteristics of the sample, their outcomes, and the methodology used. This report has virtually no data in it and sets a new low in federal grant reporting.

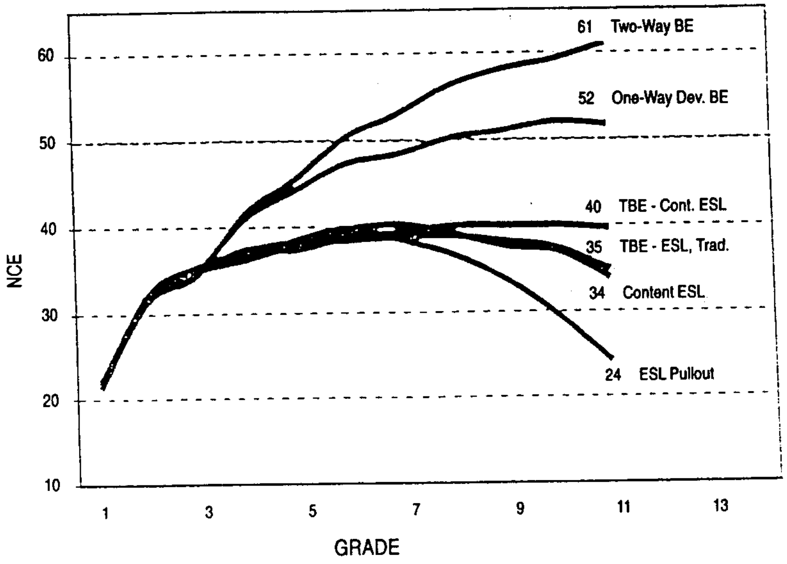
NOTES

1. I say "so-called Collier study" because in reality the study is co-authored with Wayne Thomas, a fact which is usually absent from the public discourse on the topic.
2. There is probably no issue on which people are more confused than the issue of what standardized achievement tests mean. These tests are constructed and the questions selected in order to produce a bell shaped curve wherein students are rank ordered into one hundred different categories. By definition, half the students will score at or below the 50th percentile no matter how good a job the U.S. educational system (the norming group) has done in educating children and no matter how intelligent and knowledgeable the students are. Similarly, 30 percent of the students will score at the 30th percentile or below no matter how smart the population as a whole is. These tests are useful, however, for comparing alternative programs since the limitations of the tests—that is that they only rank order students—is a constant across programs.
3. In their 1995 press release, Thomas and Collier report that they have acquired the data-set and, *where* appropriate, *combined* the results with those from their school district sites.

4. This is calculated by dividing the number of students with pre-test data in the Kindergarten through first grade and first through third grades analyses combined for each program type (Ramirez et al. [1991], Table 13, p. 60 and Table 14, p. 61 in Volume II) by the number of students who were initially in the study in each program type (Table 11, p. 51 in Volume I).
5. One of the many issues that people are confused about is what grade level means. It has no absolute meaning. It simply means the fiftieth percentile for that grade, and it is a mathematical definition that only one-half the student population can be at grade level.
6. Unfortunately, this is a technique that is used widely by school districts, and it is one of the many ways in which they are their own worst enemy. For example, school districts are constantly disseminating trends in achievement data that show declining achievement. Since each year consists of different students (perhaps one-half to two-thirds are the same from one year to the next), it is possible for each student to have an increase in achievement, but the overall average to go down because the top scoring students have left the school system. These data are then used by the press and elected officials to criticize school districts for doing a poor job of educating students when in fact the exact opposite may be true.

APPENDIX 1

Averages from Thomas and Collier's Figure 6 Produced by Hypothetical Individual Cohort Data that have Reverse Trends: Elementary Programs



APPENDIX 2

An Example of What is Wrong with a Cohort Analysis

In Table 5, I demonstrate mathematically why a cohort analysis is unacceptable. I have taken the approximate averages from Figure 1 (Figure 6 of Thomas and Collier) and created hypothetical data that produces these averages. I maintained Thomas and Collier's parameters (p. 54):

- A minimum of four years of data;
- First tested in fourth grade;
- At least one long-term cohort "confirming the general longitudinal pattern"; and
- One type of program during the elementary years.

I also maintained the overall average for each program in each grade.

The overall averages for each program in each grade, produced by the hypothetical cohorts I created, are reproduced as a graph in Appendix One so the reader can confirm that these averages do indeed produce trends in achievement that are virtually identical to those in Thomas and Collier's graph in Figure 1. This visual comparison of the graphs is necessary because Thomas and Collier present no data whatsoever in their report. The only thing we have to go on are the lines in the charts. As one can see, the two graphs look the same because I have closely approximated the averages for each program.

The first program listed in Table 5 is the two-way bilingual education program. I have created thirteen hypothetical cohort records of achievement of equal weight with a minimum of four years. Since these students were apparently first tested in grade four, there are no data for grades one to three and the average is simply entered at the bottom with identical numbers for all the programs—22, 32, 35—to approximate the curve in their graph. The first cohort I have created has scores for seventh to eleventh grade, the second cohort for sixth to eleventh grade, the third cohort for fourth to eighth grade, and so forth. The column marked "Change, 4th to last" is calculated for each individual cohort from their first score to their last score.

The average for each grade is simply the sum of the student achievement scores for that grade, regardless of the fact that each grade is composed of different students at each point in time. This is what a cohort analysis is. To illustrate for fourth grade—the average of 42 at the bottom is the sum of 20, 45, 55, 25, 10, 65, 57, 55, and 42 divided by nine (the number of students with scores for that grade). This is computed in the same way for each grade. From fourth to

Table 5
Trends in Average Achievement at Each Grade from Thomas and Collier's Figure 6 Produced by Hypothetical Individual Cohort Achievement Data with Reverse Trends: Elementary Programs

Programs	Grades											Change	
	1	2	3	4	5	6	7	8	9	10	11	4th-Last	6th-Last
Two-way Bilingual Education													
Cohort 1							75	62	66	62	61	-14	-14
Cohort 2						85	85	69	64	65	61	-24	-24
Cohort 3				20	17	16	20	18				-2	2
Cohort 4				45	45	40	41					-4	1
Cohort 5				55	56	45	51					-4	6
Cohort 6				25	24	23	22					-3	-1
Cohort 7				10	6	6	8					-2	2
Cohort 8				65	66	65	55					-10	-10
Cohort 9				57	53	55	55	51	50			-7	-5
Cohort 10						85	85	74	58			-27	-27
Cohort 11					90	75	80	62	61	60		-30	-15
Cohort 12				55	52	59	60	59	50	50		-5	-9
Representative Cohort → Cohort 13				42	45	50	53	56	58	59	61	19	11
Possible Scores of Students not Tested			*3*	*5*	*3*	*6*	*7*	*10*	*11*	*13*	*16*	*14*	*13*
Average	**22**	**32**	**35**	**42**	**45**	**50**	**53**	**56**	**58**	**59**	**61**	**19**	**11**
Avg. of 1–12	**22**	**32**	**35**	**42**	**45**	**50**	**53**	**56**	**58**	**59**	**61**	**19**	**11**
Developmental BE													
Cohort 1				18	15	15	14					-4	-1
Cohort 2				45	40	35	35					-10	0
Cohort 3							70	55	51	52	52	-18	-18
Cohort 4				60	54	53	52					-8	-1
Cohort 5					48	41	40	36				-12	-5
Cohort 6						78	72	60	51	52	51	-27	-27
Cohort 7					64	62	56	50				-14	-12
Representative Cohort → Cohort 8				41	44	47	48	50	51	52	52	11	5
Possible Scores of Students not Tested			*3*	*5*	*3*	*6*	*7*	*10*	*11*	*13*	*14*	*12*	*13*
Average	22	32	35	41	44	47	48	50	51	52	52	11	4
Avg. of 1–7	**22**	**32**	**35**	**41**	**44**	**47**	**48**	**50**	**51**	**52**	**52**	**11**	**4**

Table 5 (continued)

Programs						Grades						Change	
	1	2	3	4	5	6	7	8	9	10	11	4th-Last	6th-Last
TBE + Content ESL													
Cohort 1				18	18	18	18					0	0
Cohort 2				30	29	28	30					0	2
Cohort 3				38	33	34	35					-3	1
Cohort 4				54	50	46	46					-8	0
Cohort 5				47	48	46	42					-5	-4
Cohort 6							56	45	39	40	39	-17	-17
Cohort 7					34	33	32	30				-4	-3
Cohort 8						56	46	45	41	40	40	-16	-16
Cohort 9					50	49	49	40				-10	-9
Representative Cohort → Cohort 10					37	39	39	40	40	40	40	3	1
Possible Scores of Students not Tested			*3*	*5*	*3*	*6*	*7*	*10*	*11*	*13*	*14*	*12*	*13*
Average	**22**	**32**	**35**	**37**	**37**	**39**	**39**	**40**	**40**	**40**	**40**	**3**	**0**
Avg. of 1–9	**22**	**32**	**35**	**37**	**37**	**39**	**39**	**40**	**40**	**40**	**40**	**3**	**0**
TBE + ESL Traditional													
Cohort 1				25	24	23	25					0	2
Cohort 2				33	33	33	33					0	0
Cohort 3				43	43	43	43					0	0
Cohort 4				42	42	42	42					0	0
Cohort 5					45	45	46	45				0	0
Cohort 6				39	39	39	39					0	0
Cohort 7						43	44	41	43			0	0
Cohort 8						36	37	35	36			0	0
Cohort 9							40	38	38	40		0	0
Cohort 10				42	42	42	42					0	0
Cohort 11							35	35	36	35	35	0	0
Representative Cohort → Cohort 12				37	38	38	39	39	38	37	35	-2	-3
Possible Scores of Students not Tested			*3*	*5*	*3*	*6*	*7*	*10*	*11*	*13*	*14*	*12*	*13*
Average	**22**	**32**	**35**	**37**	**38**	**38**	**39**	**39**	**38**	**37**	**35**	**-2**	**-3**
Avg. of 1–11	**22**	**32**	**35**	**37**	**38**	**38**	**39**	**39**	**39**	**37**	**35**	**-2**	**-3**

Table 5 (continued)

Programs						Grades						Change	
	1	2	3	4	5	6	7	8	9	10	11	4th-Last	6th-Last
ESL through academic content													
Cohort 1				34	36	39	39					5	0
Cohort 2				34	36	39	39					5	0
Cohort 3				40	42	50	51					11	1
Cohort 4						28	30	32	40			12	12
Cohort 5							30	31	32	34	34	4	34
Cohort 6							40	40	40	43		3	43
Cohort 7								25	29	34	34	9	34
Cohort 8						40	46	47	47			7	7
Cohort 9				37	38	42	47	60				23	18
Representative Cohort → Cohort 10				36	38	40	40	39	38	37	34	-2	-6
Average	**22**	**32**	**35**	**36**	**38**	**40**	**40**	**39**	**38**	**37**	**34**	**-2**	**-6**
Avg. of 1–9	**22**	**32**	**35**	**36**	**38**	**40**	**40**	**39**	**38**	**37**	**34**	**-2**	**-6**
ESL Pullout taught traditionally													
Cohort 1				39	40	41	42					3	1
Cohort 2				30	42	44	45					15	1
Cohort 3				40	40	42	49	75				35	33
Cohort 4							45	48	50	54		9	9
Cohort 5					32	32	42	55	59			27	27
Cohort 6								18	25	26	29	11	11
Cohort 7							8	14	17	19	22	14	14
Cohort 8								10	17	19	22	12	12
Representative Cohort → Cohort 9				36	39	40	39	37	34	30	24	-12	-16
Average	**22**	**32**	**35**	**36**	**39**	**40**	**39**	**37**	**34**	**30**	**24**	**-12**	**-16**
Avg. of 1–8	**22**	**32**	**35**	**36**	**39**	**40**	**39**	**37**	**34**	**29**	**24**	**-12**	**-16**

eleventh grade the students in the two-way bilingual education program have on average a remarkable 19 point gain—approximately the same gain shown in Thomas and Collier's chart, Figure 1 of this article. What is interesting about this is that, although the average for each grade shows a gain over time, every single student in the bilingual program had a decline in achievement from the time they were first tested to the time they were last tested (except cohort thirteen which was inserted to maintain the rule that at least one cohort must demonstrate the underlying or average trend).

In fact, the so-called "representative" cohort has no impact on the average. If I remove cohort thirteen, the average for the program at each grade is exactly the same. This is shown in the table in the row "average of 1–12." I put this row in to demonstrate that Thomas and Collier's concern for identifying a cohort that was representative of the trend for the program means nothing because the representative cohort has no effect on the average. There is no one who would declare a program a success that had all (or 92 percent) of its students with achievement declines. Yet, the averages for the program "demonstrate" it is a success, with or without the one cohort that shows the same trend as the average.

There is a final line for the bilingual education programs consisting of hypothetical scores of students not tested, which is not calculated as part of the average for the program. I have inserted the probable scores of non-test-takers to give the reader an idea of what these averages might look like if the low-scoring untested bilingual education students were included in the computation. This row is not inserted for the ESL programs because almost all of these students are tested.

The next program in Table 5 is developmental bilingual education. There are eight cohorts with an average gain from fourth to eleventh grade of eleven points. Not quite the "gain" of the two-way bilingual programs, but still impressive. Moreover, by the seventh grade the students are supposedly scoring above the national norm. Yet, as with the two-way bilingual education program, every cohort but one (the one inserted because it is identical to the underlying trend) shows a large decline in achievement over the course of the program.

The next program in Table 5 is "TBE + Content ESL"—transitional bilingual education with the English language taught through subject matter taught in English (contrary to bilingual education theory). Again the individual cohort data is the reverse of the average for the whole program at each grade. Two cohorts showed no change from the first to the last test, seven cohorts showed a decline, and only one cohort showed a gain. Despite the fact that 90 percent of the cohorts showed no change or a decline, the average across all cohorts is a small gain.

"TBE + ESL Traditional" is the next program in Table 5. Knowing how hard it is to determine what a program actually is without going into the classroom, I cannot imagine how Thomas and Collier were able to divide TBE programs into two types—those that had ESL pullout and those that had "content ESL." As is typical of this report, they do not tell us.

Although only one of the cohorts in this program had a decline in achievement and 92 percent showed no change, the average for the program is a decline. Again, the average is representative only of the one cohort that demonstrated the underlying trend and, of course, the averages are the same with or without the representative cohort.

The last column in Table 5 shows the decline from sixth grade on for these cohorts and for the whole program. Again, although none of the individual cohorts showed a decline, across all cohorts there was a three-point decline from the sixth grade.

The next program in Table 5 is ESL through academic content. Every cohort in this program, except the representative cohort, showed a gain from the first test to the last test. Yet, the program shows a small decline from fourth to eleventh grade. Similarly, every cohort showed a gain, some of them quite large, from sixth grade to the last test, but across all cohorts there was a decline of six points from the sixth grade.

The final program in Table 5 is ESL pullout taught traditionally. Every cohort, except the representative cohort, showed a gain, some of them quite large from fourth grade on. Yet, the program as a whole is considered a failure—because it has a twelve point decline. From sixth grade on every cohort showed a gain, some of them quite large; however, the program as a whole had a sixteen point decline!

In short, Table 5 demonstrates how overall averages for each grade can be produced by shorter individual cohorts with exactly the opposite trends. It is for this reason that a simple descriptive cohort analysis is not a valid means of evaluating programs and will continue to be rejected by scientists.

References

Boston Public Schools. (1994). School profiles 1992–93. Boston: Office of Planning, Research, and Development.

California Department of Education. (1997). Language census report for California public schools. Available from CDE web address: http://www.cde.ca.gov.

Cazabon, M., Lambert, W.E., and Hall, G. (1993). Two-way bilingual education: A progress report on the Amigos Program. Santa Cruz, CA: National Center for Research on Cultural Diversity and Second Language Learning.

Collier, V.P. (1995). Acquiring second language for school. Washington, DC: National Clearinghouse for Bilingual Education.

Lambert, W.E., and Cazabon, M. (1994). Students' views of the Amigos Program. Santa Cruz, CA: National Center for Research on Cultural Diversity and Second Language Learning.

Lindholm, K. (1998). Declaration, in the case of *Valeria G., et al. v. Pete Wilson, et al.*, June 2.

Ramirez, J.D., Pasta, D.J., Yuen, S.D., Billings, D.K., Ramey, D.R. (1991). Final report: Longitudinal study of structured immersion strategy, early-exit and late-exit transitional bilingual education programs for language-minority children. Report to the U.S. Department of Education, Washington, DC. San Mateo, CA: Aguirre International..

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

———. (1996b). *Bilingual education in Massachusetts: The emperor has no clothes.* Boston: Pioneer Institute.

San Jose Unified School District School Profiles. (1998). School profile for 1997–98 [with data from 1996–97]. San Jose, CA: San Jose Unified School District.

Steel, L., Levine, R., Rossell, C., and Armor, D. (1993). Magnet schools and issues of desegregation, quality and choice, phase I: The national survey and in-depth study of selected districts. A report to the U.S. Department of Education.

Thomas, W.P. and Collier, V. (1995). Research summary of study in progress: Results as of September, 1995, language minority student achievement and program effectiveness.

———. (1997). School effectiveness for language minority students. *NCBE Resource Collection Series*, No. 9. George Washington University. Downloaded from NCBE web address: http://www.ncbe.gwu.edu/ncbepubs/resource/effectiveness/index.html.

Zacaria, R. (1998). Clarification of English academic testing results for Spanish-speaking LEP fifth graders. Report to the Los Angeles Board of Education, March 4.