Disordered Data and Murky Models

Critique of Wayne P. Thomas and Virginia P. Collier, "A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement," Center for Research on Education, Diversity and Excellence, 2002.

By Dr. Christine H. Rossell

Lexington Institute

July 2008

Executive Summary

Few education research studies have garnered greater attention in education-policy circles than the analyses of programs for language minority students conducted by Virginia Collier and her George Mason University colleague Wayne Thomas. In fact, much of the public attention received by their most recent paper occurred *before* the report was self-published in 2002. Since then, the paper has been cited and its findings discussed in public documents and proceedings around the country, particularly to support teaching English learners in their non-English native language.

This critique, by Boston University professor of political science Christine H. Rossell, offers the first systematic review of the 2002 "National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement," and its findings. Among its conclusions:

- Numerous assertions of educational benefits remain unsubstantiated, or worse, influenced by significant factors ignored by the authors.

- While Thomas and Collier repeatedly refer to various assertions that "the research to date has found," in fact, there is no research other than theirs that has found this and their research often does not withstand close scrutiny using standards of social science research.

- The research design contains fundamental flaws, such as the absence of a control group — similar students in alternative program(s) — and the absence of statistical control (if there is no random assignment) for other variables that explain achievement, render many significant conclusions highly questionable.

Details follow.

Disordered Data and Murky Models

By Dr. Christine H. Rossell

July 2008

Introduction

In bilingual education circles, Virginia Collier has been a media star ever since she embarked on a nationwide tour in 1999 promoting the results of her as-yet unreleased study with Wayne Thomas. At public meetings around the U.S., she handed out a five-page summary of their preliminary results consisting of two pages of text, two pages of line graphs, and a one-page list of program definitions (Thomas and Collier, 1995). In no time, the "Collier Study" had become another factoid in the controversy over bilingual education. Even though no one had actually read it – since it had not been written – the report was being cited everywhere from Arizona to Florida as proof that bilingual education, particularly two-way bilingual education, was superior to all other programs for limited English proficient (LEP) children.

Two years after the media blitz, the full study was released (Thomas and Collier, 1997). The report was prepared with funding from the Center for Research on Education Diversity and Excellence, a national research center funded by the U.S. Department of Education. Although 96 pages, it contained no more data than the two pages of line graphs in the summary handed out two years earlier. Even stranger, the lines, which looked hand drawn, had not changed at all although the sample size had changed since the preliminary report. The study was massive including "over 700,000 language minority student records, collected by the five participating school systems between 1982 and 1996..." (Thomas and Collier, 1997: p. 30).

No data was presented in the 1997 report beyond the simple descriptive line graphs, no publication since then has presented the missing data from that time period. Their later study in 2002, which is the topic of this report, has plenty of data, but none of it is from the 1982 to 1996 time period covered in their 1997 report. So that study is still a mystery (see Rossell, 1998).

Although all four characteristics of a scientific study are essential, they are not found in the case studies presented by Thomas and Collier.

What is a Scientific Study?

It is important for the reader to understand what a scientific study is to understand issues with how Thomas and Collier apply those principles. The criteria for a scientific study (see Rossell and Baker, 1996a, 1996b; Rossell, 1998) 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 (also called a control group) – for example, similar LEP students in alternative 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 statistically in order to give each group a level playing field. 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, they are not found in the case studies presented in Thomas and Collier, 2002.

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 another program, a negative effect if they would have scored at the 40th percentile in another program, or no effect at all if they would have scored at the same level in another program. It is only this comparison of students in one program to students in another program that enables us to evaluate what a score at the 30th percentile means.

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. Random assignment is, however, rarely possible and is not used in any of the sites studied by Thomas and Collier. Rather, students were assigned to, or selected themselves for, different programs based on their individual characteristics such as motivation, intelligence, social class, or learning problems. The differences that existed before the program that caused the student to select herself or be assigned to a 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, as well as other important characteristics.

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. Thomas and Collier do not follow these rules or perhaps it is their district "collaborators" who do not follow the rules.

Thomas and Collier, 2002, claims to pick up where Thomas and Collier, 1997 left off. The time period is now 1996-2001. However, they still have five sites, but have added two sites in rural Maine. Simple math suggests that they have lost two sites, but they do not acknowledge these. They also seem to have lost a lot of data since the total number of student records is now only 210,054, down from over 700,000. The five sites in this version of their research are two school districts in the rural St. John Valley of Maine: Madawaska School District and Maine School Administrative District #24; the Houston, Texas Independent School District; Grant Community School in Salem, Oregon; and a school district in the Southeast that wishes to remain anonymous. The sites were apparently selected for geographical representation rather than the facilitation of a valid, scientific analysis.

St. John Valley, Maine

In this first section, Thomas and Collier study two tiny school districts in rural Maine. The Madawaska School District is a district with two schools, an elementary school with a PK-5 French Immersion program where French is the second language that is learned and a middle/high school (6-12). The entire school district has 753 students – a decline from the number cited by Thomas and Collier for 1997. The second district, Maine School Administrative District (MSAD) #24, similarly has two schools – an elementary school with about 267 students and a 7-12 high school with about 222 students for a total of 489[1] in 2001 – again a decline from the total cited by Thomas and Collier for 1997.

As noted by Thomas and Collier, the students start school fluent in English: "very few students used the French language to any significant degree in the home or community, due to the high level of linguistic assimilation within the community because past generations did not have the opportunity to be schooled in French" (p. 51).

> *Why would a researcher choose a community and program so different from bilingual education as it is commonly practiced?*

Clearly, this program is *not* bilingual education as it is commonly practiced in the U.S. because if it were, the students would begin school limited in English and would study French in order to enhance their cognitive abilities and knowledge. Since these students were already fluent in English, a fact which Thomas and Collier spend considerable time discussing (pp. 51-53), the French instruction is actually French immersion or French enrichment. By the middle school years, the amount of French instruction is only about an hour a day.

The fact that French is a "heritage" language, a point made much of by Thomas and Collier, strikes me as irrelevant. If the children in this program have to learn French, it is a foreign language for the students in this program regardless of their "heritage." Why a researcher would choose a community and program so different from bilingual education as it is commonly practiced – LEP students learning to read and write in the native tongue, getting subject matter in the native tongue, and receiving increasing amounts of English

[1] Source: http://nces.ed.gov/ccd/pubschuniv.asp.

instruction – is baffling. In addition, there are many Spanish immersion programs in the U.S. Why select a unique French immersion program in rural Maine for fluent English speakers? Even if Thomas and Collier wanted the Northeast represented, there are a number of Spanish-English two-way immersion programs in Massachusetts.

The statistics used to study this program, itself of dubious relevance, are flawed. Having claimed in the research methodology section that each cohort was studied separately, the authors reverse themselves in the analysis – "these first four data displays do not follow precisely the same group of students across time and therefore they are labeled cross-sectional" (p. 62). In Figure A-1 and A-2 of their report (pp. 77-78), they show the former LEPs in what they call "bilingual immersion," and the English mainstream classroom. They claim that the two groups had exactly the same average Terra Nova English reading and math scores at the beginning of each program.[2] This is puzzling since the parents who volunteer their children for a French enrichment program are likely to be different from those who did not.

Achievement in math is quite erratic for the bilingual immersion students – a sharp increase of 6 points in the second year, but a sharp decline of 6 points in the third year. The language achievement shows a sharp increase of 13 points in the second year. These fluctuations are caused by two problems. First, the number of students in the first, second, third, and fourth years of the French bilingual immersion program varies considerably from year to year (from 41 to 90 to 67 to 54) because it does not represent the same students over time. Second, the French bilingual program represents only 12 percent of the enrollment of the schools they are in, thus giving further credence to the notion that this is a small, elite program.

In fact, if we look at the data in Table A-6 (p. 91) of the Thomas and Collier report, we see that the 1997 cohort in the bilingual immersion program (French enrichment) began with Terra Nova scores in 1st grade in 1997 that were 12 points higher for reading, 7 points higher for language, and 6 points higher for math than the non-immersion students, not the 0 difference shown in Figure A-1. The 1998, 1999, and 2000 cross-sectional cohorts in the French bilingual immersion program had similar advantages when they started school.

Although the cohorts are of different lengths,[3] they are combined into the same cohort based on number of years in program or grade in program and LEP status. As I demonstrated in Table 5 of Rossell (1998), the problem with cohort analysis of this type is that it is possible to have all students with declines in achievement, but because of the changing composition of the cohort (as students come and go from a school or program), the overall average can go up. The reverse is also true. This is commonly called Simpson's

[2] Thomas and Collier report all scores in NCE, which stands for normal curve equivalents. It is the area under the normal curve that a percentile represents. These scores are created by a formula from national percentile scores and both have a mean of 50. Only three scores are exactly the same on the two measures: 1, 50, and 99. NCE scores are superior to percentile scores for statistical analysis because they have a normal curve distribution. Percentiles, however, are easier to interpret for most people.

[3] The 1997 cross-sectional cohort contains only two years of data starting in 1st grade, the 1998 cohort contains six years of data starting in 1st grade, the 1999 cohort contains seven years of data starting in 1st grade, and the 2000 cohort contains six years of data starting in 2nd grade.

Paradox – the reversal of the direction of a comparison or an association when data from several groups are combined to form a single group (Bracey, 2003; Moore and McCabe, 2002; Intuitor.Com, date unknown).

The analyses in Tables A-6 to A-11 focus on change in the scores of both program types (bilingual immersion and non-bilingual immersion). The trends don't make any sense nor do they agree with the positive trends shown in Figures A-1 and A-2. Most changes show a *decline* in reading, math, and language, some as large as 17 points. If we were to take these data at face value, the conclusion would be that there is a penalty to being taught in French part of the day.

Thomas and Collier briefly try to explain the decline in achievement, but ignore it elsewhere in the report. They do, however, caution the reader that the numbers may not be reliable due to the small sample size for the bilingual immersion students. They further explain that the students in this school and in this program were above average and such students cannot be expected to continue to increase their achievement. Both statements are valid, raising questions about the relevance of their analysis.

To get an idea of how unique this program is, we can look at the scores of students in other highly regarded "two-way immersion" programs. As shown in Table 1 of this report, the St. John Valley, Maine French enrichment students' test scores are closer to those of the Anglo (white) students in two highly regarded two-way immersion programs – the River Glen School in San Jose, California and the Hernandez School in Boston – not to the Hispanic students. For normal curve equivalents (NCE), 50 is grade level, shown as the score in parentheses in the top row of that table. The St. John Valley students are above grade level, and so are the Anglo students in the San Jose and Boston two-way immersion programs. These Hispanic students' scores are below grade level in both the River Glen and the Hernandez schools.

The St. John Valley French enrichment students are also closer to the Anglo students in another highly regarded two-way immersion program – Amigos – in Cambridge, Massachusetts, shown in the lower part of Table 1. The scores for this program are in grade equivalents, not NCEs, but if we compare the scores of students in the Amigos program to grade level (the numbers in parentheses in the top row in the lower portion of the table), the Anglo students are well above grade level and the Hispanic students are below grade level by third grade. In other words, the two French immersion programs that Thomas and Collier study are statistically elite programs consisting *only* of the higher scoring Anglo students found in two-way immersion programs in the U.S.

Thomas and Collier claim their analyses:

> …dramatically demonstrate that students schooled through two languages outperform those schooled through one language. These bilingually schooled students have also acquired French at no cost to their English achievement (Thomas and Collier, 2002, p. 68).

Table 1

Reading Scores (NCEs) in French Immersion programs in St. John Valley, Maine and Reading Scores in Highly Regarded Two-Way Immersion Programs

	GRADE IN SCHOOL							
	1	2	3	4	5	6	7	8
GRADE LEVEL (NCE)				(50)	(50)	(50)	(50)	
French Immersion-ME[1]				**66**	**64**	**63**	**58**	
River Glen School-San Jose[2]								
Whites					**58**			
Hispanics					38			
Hernandez School-Boston[3]								
Whites	-----	**75**	**93**	-----	-----	-----	-----	-----
Hispanics	-----	-----	-----	46	39	39	36	44
GRADE LEVEL (GE)	(1.7)	(2.7)	(3.7)	Grade Equivalent Scores				
Amigos Program, Cambridge[4]								
Whites	**1.3**	**5.1**	**4.7**	Grade Equivalent Scores				
Hispanics	1.3	3.1	2.9	Grade Equivalent Scores				

Key: ----- means less than seven students took the test.

[1] Source: Thomas and Collier, 2002, Table A-6, p. 93.

[2] Source: San Jose Unified School District (1998) (percentiles converted to NCEs)

[3] Source: Boston Public Schools (1994) (percentiles converted to NCEs)

[4]Source: Cazabon, Lambert, and Hall, (1991), p. 8, 14-16.

Their analysis of variance[4] of reading, language, and math scores, however, showed no difference between the achievement of students in the mainstream classroom and those in the French enrichment program. Thomas and Collier also present a hierarchical stepwise regression analysis in Table A-18 (p. 115),[5] but only use the change in r^2**,** which is the change in the variation in reading achievement, to assess the importance of two variables: years in the bilingual immersion program and socio-economic status. They do not even look at math achievement.

There is no b coefficient and no Beta and thus no sign next to the two independent variables: years in the bilingual classes and socio-economic status. Nor is there a significance level. It is possible that the number of years in the bilingual program has a *negative* effect on one's reading achievement. Thomas and Collier do not present this essential data and a negative effect cannot be determined merely from the data they do present – the r^2 or significance of the change in r^2.

[4] This statistical analysis compares the means (on some measure) of each group to the variation (on some measure) within each group. The difference between groups will be statistically significant, when the mean between groups is large enough and the variation within groups is small enough, also taking into account sample size.

[5] Stepwise is not a recommended regression approach because it allows a computer make the decision as to whether a variable should enter the equation.

Furthermore, basing the importance of a variable on the increment in r^2 is an approach that is not only decades out of date (the statistics book they quote was published in 1975), but raises suspicion that they are trying to hide something since later in their report, they do use multiple regression with b coefficients, Betas, standard errors, t values, and significance levels that allow the reader to assess the direction of a relationship and whether it is statistically significant.

In short, not only is the French bilingual immersion program irrelevant to the policy debate over the best way to educate immigrant LEP children, but their statistical analysis is invalid and can tell us nothing about whether students had higher English achievement because they were in the French bilingual immersion program or if they started with higher achievement, which is suggested by the descriptive data in their tables.

Houston, Texas Independent School District

Although this school district would presumably have real bilingual education, the analysis is marred by the same problem as the St. John Valley, Maine analysis.[6] The cohorts are "cross-sectional," that is different students in different grades in the same year.

To describe these students as having scores that "go down" throughout their schooling is inaccurate and misleading.

Once again, Thomas and Collier throw caution to the wind. They conclude that with regard to English language learners who received all of their schooling in English – "beginning in 9th grade, their scores began to drop, and they reached only the 40th NCE by 11th grade." However, Thomas and Collier are wrong. One can conclude *nothing* about the relative effectiveness of a program from a table of the reading (Table C-1, p. 155) and math scores (Table C-2, p. 156) of *different students* in *different grades* in the same year (1999). The students who are in 11th grade in 1999 are *not* the same students who are in grades 2-5 in 1999 and Thomas and Collier should know better than to discuss them as if they are.

Relying on the same data in one year (1999), they conclude that:

> … the biggest shock is the achievement levels of those students who were not proficient in English upon enrollment in the Houston schools whose parents signed a waiver requesting that their children be placed in the mainstream, with no bilingual or ESL support…these students were doing very well when first tested…in second grade. Their scores lowered to the 45th NCE (40th percentile) by third grade and continued to go down throughout their schooling… (pp. 128-129).

[6] Houston, which follows the St. John Valley section, labeled A, is labeled C, without explanation.

To describe these students as having scores that "go down" throughout their schooling is inaccurate and misleading. The accurate way to describe these cross-sectional cohort data is to say that former LEP students in the 11th grade whose parents refused services in 1999 have lower scores than former (and different) LEP students in the 2nd grade whose parents refused services in 1999. The reason for this is not known. Perhaps elementary students whose parents refuse services have children whom they believe to be academically able – indeed they are scoring almost at grade level in 1999 as shown in Tables C-1 and C-2 (pp. 155-156). High school students, however, play a large role in their own academic programs and it may be that high school students who refused academic services in 1999 are students with little academic ambition or ability. Regardless, different students in different grades in the same year can*not* have scores that go down throughout their schooling as Thomas and Collier assert since there is only one point in time.

Thomas and Collier also examine the Spanish scores in reading and math of native Spanish speaking students who were in bilingual education and compare them to the English scores in reading and math of native English speakers in the mainstream classroom in 1999 in grades 1- 8. Not only are these different students in each grade, but comparing the English scores of native English speakers to the Spanish scores of native Spanish speakers is comparing apples to oranges. If these tests are comparable (and it is unlikely they are), the data show that the students in the bilingual education are more academically able than the native English speakers in the mainstream classroom to begin with, which would invalidate their simple descriptive comparisons and conclusions.

On p. 131, having only presented simple descriptive data of the scores of different students in different programs at different grade levels in one year, Thomas and Collier boldly assert that "... it is very evident in the district-wide data that bilingually schooled students outperform monolingually schooled students..." Such conclusions drawn from the data they present are both inaccurate and misleading. Furthermore, throughout the report they make statements that "the research to date has found" one conclusion or another when, in fact, there is usually no "research" other than theirs that has found this and their "research" is simply not scientific.

Although there are only a small number of native English speakers in the two-way program, only a minority of whom know enough Spanish in any grade to take an achievement test in Spanish (pp. 158-160 and 190-192), Thomas and Collier claim that this small number of native English speakers has a positive effect on the *Spanish* scores of native Spanish speakers.

Thomas and Collier also examine Spanish achievement outcomes by number of years in the program in simple, descriptive tables (C-5, C-6, pp. 161-189). To no one's surprise, the more years in the program, the better your Spanish is. But, is there anyone who ever doubted that being taught in Spanish improves your Spanish? What is controversial is the theory that being taught in Spanish improves your *English*.

Before I would recommend a two-way Spanish immersion program to native English speakers, I would want to see a statistical analysis of the English achievement outcomes of native English speakers in a two-way immersion program and the same sorts of students in a mainstream classroom (i.e. controlling for the socio-economic status, parent's education and test scores of the native English speakers before they started the program). Thomas and Collier do claim to have a control for the socio-economic status and neighborhood of the *school* in Table C-7 (p. 190), but this cannot replace the socio-economic status of the *student* since these students *volunteer* for all programs except the transitional bilingual education program, which according to them is "not successful."

Throughout their analysis of Houston, Thomas and Collier continue to inappropriately discuss the patterns in different grades in a single year as if they are patterns across time of the same students, which they are not. The report also makes such assertions as:

> ... schooling intensely through Spanish in the early grades seems to enhance their English achievement, when compared to district-wide test scores, which cluster around the 50th percentile. This provides still more evidence that in the long-term, bilingually schooled students outperform their monolingually schooled students (p. 139).

The implication that the two-way immersion program did this is unsubstantiated. Imagine that the kind of parents who voluntarily decide to enroll their children in a two-way immersion program tend to have children who score in the 80th percentile and above. If this is the case, the two-way program has *harmed* these native English speakers since the scores Thomas and Collier report tend to be in the high 50s and low 60s. Since Thomas and Collier do no statistical analysis at all, we do not know the effect of the two-way program or any other program on the achievement of the students in them.

What is controversial is the theory that being taught in Spanish improves your English.

With regard to *LEP* achievement, the analysis they should have done to test the effectiveness of the two-way program would be to compare the English achievement scores of LEP students in different programs, statistically controlling for the other characteristics of the students that would also affect achievement such as parents' education, profession, the student's free or reduced lunch status, the school characteristics, and ideally (although rarely available) pre-program test scores to determine the relative impact of the different programs on English achievement.

With regard to the *native English speakers*, the analysis they should have done to test the effectiveness of the two-way program would be to compare the English achievement scores of native English speakers in different programs, controlling for the variables in the above paragraph. Since they did not do this, their conclusions on pp. 139-142 are unsubstantiated.

Grant Community School, Salem, Oregon

Section D of the Thomas and Collier report examines a two-way immersion school, Grant Community School, in Salem, Oregon. According to Thomas and Collier, all students enrolled in the school are in the program. However, according to the school administration, the school has a neighborhood attendance zone and the students in the two-way program are volunteers from all over the school district. The neighborhood attendance zone students may or may not be in the two-way immersion program. Therefore, there is a self-selection bias problem with this program that cannot be corrected since Thomas and Collier have no control group of similar students in alternative programs.

Figure D-1 (p. 230) purports to show progress over time in the learning of English and Spanish on the SOLOM, an oral examination that is administered in Spanish or English. This is an inappropriate test to use since the Spanish speakers start off scoring at 25, the highest level in Spanish and the English speakers start off scoring at 25, the highest level in English so there is no room for improvement in those two groups in their native tongue.

Of the four trend lines in Figure D-1 (one seems to be missing) drawn from Table D-1 (none missing), only one trend line is important to the debate over bilingual education and that is the line for Spanish speakers in English. That line suggests that their oral English improves over time as does their oral Spanish, although only a bit since they started at the top. However, there is no control group. The interpretation of positive trends as positive impacts when there is no control group is, once again, an error. If native Spanish speakers in a mainstream classroom with English as a Second Language (ESL) pullout are making greater progress in English on this exam than the students in the two-way program, then in fact the two-way program actually harms the students even though both are making "progress."

There is no valid control or comparison group.

Despite all the numbers presented in the figures and tables for this section, it is in the end an anecdotal case study, relying on discussions with teachers about whether 90-10 or 50-50 is better, musings about the principal's "strong" Native American ancestry – Chickasaw, Chiricahua Apache, and Cherokee – as well as some German and Irish ancestry, which it is implied make him a good principal for a Spanish two-way bilingual immersion program. Not only is there no valid control or comparison group – LEP students in alternative programs – the only analysis compares the percentage of students in the two-way program to state standards.

When the Grant Community School has a smaller percentage of Hispanic students in 5th grade meeting the state standard than the district and the state as a whole, Thomas and Collier excuse it as a function of new students at that grade. However, when the school has a higher percentage of Hispanic students in 3rd grade meeting the state standard than the district and the state as whole, they credit it to the two-way immersion program.

A simple comparison of the achievement of students in a school to those in the district and the state (Table D-17) proves nothing. If the students in the school started off much higher than the state [as suggested by the simple, descriptive data Thomas and Collier present in Table D-17 (p. 254)], the students in the two-way program could be harmed by the program, but still be doing better than the district and the state.

In Tables D-20 and D-21, Thomas and Collier present a hierarchical stepwise regression similar to the one used to assess the French two-way immersion programs in Section A of their report. Again, they attempt to determine the relative importance of student socio-economic status (not defined) and years in the two-way program by the increment in r^2 when each variable is entered first. As noted above, this is an approach that is not only invalid, but raises the suspicion that they are trying to hide something since they present no regression coefficients, no standard errors or t values, and no significance levels that would allow the reader to predict achievement from each variable and to know the direction and significance of the relationship. However, as crude as their statistical approach is, even if they had included the above essential information, it would have told us little because an even more essential factor is missing from their data – LEP students in alternative treatments and control variables (parents education, profession, socio-economic status) that are also known to influence achievement. Even if years in the program has a positive relationship to English achievement, if the progress is less than for similar students not in the program, the effect of the program is in fact harmful to a student's English achievement.

Mid-Sized Urban Site in the Southeast

Remarkably, this is not an analysis of the effectiveness of alternative programs for LEP students because the district only had an ESL program. One can only wonder then why Thomas and Collier would pick such a school district in the first place if their interest was in determining the effect of alternative programs on the achievement of immigrant LEP children? Geographic representation is a questionable reason when the site is so clearly inappropriate.

One goal Thomas and Collier have is that LEP students be at "grade level." Thomas and Collier cite Collier, 1989; Cummins, 2000; Hakuta, Butler, and Witt, 2000; and Thomas and Collier, 1989 for the proposition that it takes 4-7 years for LEP students to reach "grade level" performance using the nation (the norming group) as the control group. However, grade level is not the appropriate standard because all standardized tests, even state proficiency tests that claim to be criterion referenced, are constructed so that half of all students who know no language other than English are below grade level (see Rossell and Baker, 1988; Rossell and Baker, 1996; Rossell, 2000). Therefore, across the U.S. only half of all LEP students will reach grade level. Furthermore, if an LEP student's true score (in the absence of a language barrier) is well above grade level, whatever program has produced grade level achievement may not have done enough.

Another goal for this school district, according to Thomas and Collier, is to close the achievement gap between three groups of students – 1) formerly LEP, 2) language minority (LM) who were never LEP, and 3) native English speakers. This too is unrealistic.

Although there are exceptions, these three *groups* will usually have unequal achievement levels on standardized tests during their school years. This occurs for two reasons – initial scores and different family environments. Assuming the native English speakers are not an exceptionally poor, at-risk group, they will have the highest test scores because they come from a family where English is spoken as determined by the home language survey that all students take when they enter school. In addition, this group will tend to have lower mobility than any immigrant group. And in fact Figure E-1 (p. 287) of the Thomas and Collier report, shows the native English speakers to have the highest scores.

The former LEP students tend to have the lowest scores in this district because LEP students are defined by their low achievement and those recently redesignated will still have fairly low scores. Those students will drag down the scores of the group as a whole which will, of course, have some very high scorers.

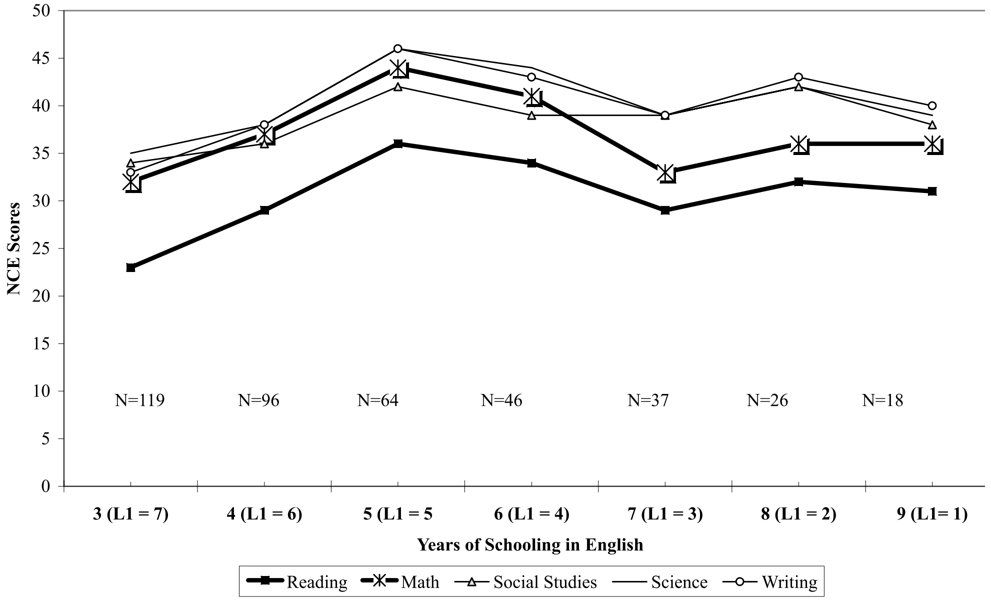
The "Language Minority but never LEP" group will tend to have higher scores than former LEP students, but lower scores than native English speakers since their home environment is non-English speaking (which is why they are designated LM). Thomas and Collier, however, assert that the "LM but never LEP" group did not catch up to the native English speakers in five years because they were educated in English rather than bilingually. The likely relationships discussed above, between home language and achievement, suggest that, unless the native English speakers in the public schools are unusually poor and at-risk, there is no reason to expect the achievement gap to be closed in any particular time period. Certainly, Thomas and Collier present no evidence that the native English speakers are unusually poor and at-risk. Therefore, one cannot conclude that whatever program is in place is responsible for the gap not being closed. It is a concern that few people who read this report will be able to interpret the many tables and most will just skip over them assuming that they do in fact support the conclusions of Thomas and Collier.

Thomas and Collier present another "quasi-longitudinal" analysis (Figures E-6 and E-7 and Tables E-6, E-7, E-8) of different students at different points in time, although there is overlap. They draw conclusions about the environment these students were in prior to coming to this district that I doubt the school district actually knows. First, they claim to know how many years the student has been in the U.S. and in what language they were educated in (apparently even when it was outside this district). Second, they claim to know how many years of schooling the students received in their primary language in their "home country." It strikes me that such data, even if the school district actually has it, is unlikely to be accurate. There are also apparent errors or contradictions in labeling that suggest this is the case.[7]

[7] Figure E-6 and Table E-6 of Thomas and Collier have the title "length of residence in the U.S" but the column heading is "years of schooling in English." Thomas and Collier claim Table E-7 is the next year after Table E-6, but that cannot be since the two sets of data are identical except that E-6 does not have one and two years residence. The only question is why? Simple math indicates that the sample size for the two years of residence in the U.S. is 106 students and if divided by two would be 53 students for the two data points. The sample size would then jump from around 53 to 119 for three years of schooling in English.

Figure 1

NCE Scores by Years of Schooling in English, SE School District



Source: Thomas and Collier, pp. 292.

Thomas and Collier present data on the relationship between achievement and the number of years students have been schooled in English (Table E-6, p.311). Their data is illustrated in Figure 1 of this report. It shows achievement over the number of years schooled in English, as well as the average sample size N for each of the four tests from their Table E-6 (p. 311). The number of students who have been schooled in English declines from 119 students for three years to 18 students for nine years. The test scores peak with five years of schooling in English when the number of students has declined from 119 to 64, then they decline, then they go up again.

We do not know the characteristics of the students who left the school district and those who stayed – a common pattern in heterogeneous schools is for the more academically able students to leave with each successive grade. But we do know that the sample size is very small (18 in the final group), and the data on home language schooling is unlikely to be accurate. Instead of concluding this is questionable data which a sensible person would not want to rely on, Thomas and Collier conclude that:

> … these findings gave further confirmation to the decision to provide some primary language (L1) content instruction for students with little or no formal schooling in L1 (p. 273).

This conclusion can hardly be substantiated by the data presented.

Thomas and Collier use data for "Years Below Grade Level Upon Entry" as a proxy for "Lost Schooling in Home Country." Even if this were real data on lost schooling in the home country, it still would not prove that the student needed to be educated in the primary language or be sent back to their home country to be educated there. All that it would show is that children who have not been in school in their home country have big problems and one can only imagine what those are – poverty, abuse, learning disabilities, etc. To determine if the remedy is more education in their primary language, one would have to statistically analyze the achievement of such students in alternative programs – bilingual education, mainstream with ESL pullout, and sheltered English immersion.

We do know that the sample size is very small (18 in the final group), and the data on home language schooling is unlikely to be accurate.

But these data are apparently *not* real and we do not know whether any of these students lost schooling in their home country. We only know that if they are below grade level in Spanish, which half the Spanish speaking population is (as is half the English speaking population in English), they are assumed to have lost schooling in their home country equal to the number of grades below grade level that they test at.

Thomas and Collier finally present a multiple regression analysis (Table E-16, p. 320) with all the data usually found in such analyses [unstandardized b coefficients, standard errors, Betas, t values, significance levels, and even confidence intervals (not usually found in such tables)]. The first question that comes to mind is why was a complete regression analysis presented only for this district? Why was only the increment in r^2 presented in the two districts in rural Maine and the Grant School in Oregon? It is not because they did not have the data, since if you have the data to present r^2, you have the data to present all of the above information. In addition, it is not a matter of limited time since the statistical programs that produce r^2 automatically produce all of the above information. There are only two possibilities that make any sense. The first is that the person who did the analysis for this section, did not do the analysis for the other sections and Thomas and Collier did not notice the difference in analyses or understand that the other analyses were deficient. A second possibility is that they selected analyses to produce desired results. At this point, we do not know.

In their first regression analysis, Thomas and Collier predict the achievement of former LEP students in the 11th grade by 8th grade reading scores, and "years of lost schooling," which is actually the years below grade level in Spanish when the student entered the district, and seven other variables. Of the nine variables entered into this equation, only two are statistically significant – 8th grade achievement and free lunch in predicting 11th grade achievement.

In a third multiple regression analysis shown in Table E-18 (p. 322), non-significant variables are thrown out by the computer and the only variables that remain are the statistically significant ones: in this case, 8th grade achievement and free lunch – in predicting 11th grade achievement.

Thomas and Collier claim that grade completed in prior schooling is the inverse of years of schooling in English, and that this shows once again that the more primary language schooling, the higher one's achievement. Yet they put them in the same equation where they claim grade completed in prior schooling "edged out" (p. 280) years of schooling in English. If the variables really were the inverse of each other, valid statistical analysis would not have permitted them to remain in the same equation as was the case in the prior regressions in Tables E-16 (p. 320) and E-17 (p. 321). Therefore, they are most definitely *not* the inverse of each other. In addition, why was this particular variable not significant in the previous equation? They have no theory for the inclusion of each of these variables, which is perhaps why they let the computer do it.

It seems highly unlikely that these variables are accurate and there is no way to check since the school district wishes to remain anonymous. The authors do not bolster one's confidence in the accuracy of these variables since they use terms interchangeably that are supposed to be different ("years of schooling in English" is also called "length of residence in the U.S." and "time lost in home schooling" is also called "years below grade level").

Finally, Thomas and Collier do not analyze math outcomes using the above regression analysis approach because they claim that reading is more important. Nevertheless, math outcomes are presented throughout the report and in this section in many simple, descriptive tables. So what are we to make of this? Math outcomes are important enough to take up many pages in simple misleading tables, but not a half page of regression analysis? One can only be suspicious of such logic.

Their repeated conclusion that "the more primary language schooling that these students had before arriving, the higher their achievement in English, in the long term" is not substantiated by their analysis There is no control group – students in alternative programs – because the district only has one program, an ESL program. A host of ambiguous and contradictory variables measuring the students' characteristics before they arrived and on entry are unlikely to be accurate and cannot substitute for a control group.

There is no control group — students in alternative programs — because the district only has one program.

Thomas and Collier do list the variables used in this section and their definitions in Appendices A and B of their report, but the definitions do not clarify at all. The number of years of interrupted schooling has no reference – what country, what district, what school – and no real definition. The same is true of age on arrival and length of residence. Are we talking about a different country, a different district, or a different school? Is it possible that the school district staff who keep these records do not consistently agree on this? Or even that schools would be able to reliably collect and track such data reported from immigrant parents with their high mobility in the U.S.?

Certainly, number of grades below grade level on entry to the district cannot substitute for years of lost schooling in the home country. Finally, variables measuring the students' experience before they came to the district and placement when they got to the district cannot substitute for a control group – LEP students in alternative programs. In short, such data seem unlikely to be accurate.

Thomas and Collier's Conclusions

The conclusions of Thomas and Collier begin with the standard platitudes about how important a high-quality program is. Their summary of quantitative analyses repeat the same errors found in the different sections. They ignore the small sample sizes, the many analyses of different students in different grades as if they were the same students over time, the fact that the many simple, descriptive tables and figures prove nothing, and the failure to present complete statistical analyses until the very last site – one that has only one program, no control group, and wishes not to associate itself with their results.

The recitation of average achievement scores for students in different programs which they do throughout the report and again in the conclusions proves nothing. If a group of students reaches the 47th percentile, the program they are in has harmed them if another program would have allowed them to reach the 57th percentile at that point in time.

Thomas and Collier also spend two pages repeating the Spanish achievement results. Again, not only are these derived from simple, descriptive analyses, but they apparently do not see any contradiction in arguing that more instruction in the native tongue improves one's native tongue (is there anyone who doubts this?), but *less* instruction in English improves one's English. They also make the same statement about native English speakers: less instruction in English for native English speakers improves their English (p. 331)! Of course, there is no statistical analysis anywhere in this report that controls for the characteristics of the students who choose these two-way bilingual programs.

Conclusion

Table 2 summarizes the most important characteristics of the sites that Thomas and Collier studied and my assessment of their statistical analysis. The first two of the five sites Thomas and Collier studied were of French-English two-way immersion programs in Maine. The students in these programs were English speakers, and French was their heritage language. In 2001, there were only 249 two-way immersion programs in the U.S. Of those, 234 were Spanish-English and only 5 were French-English (the rest were other language immersion programs). Why select a two-way immersion program to study that represents only two percent of the programs in the U.S. for not one, but two sites? So 2/5 of their sites are irrelevant to the debate in the U.S. over whether bilingual education is the best way to educate immigrant LEP children, even if their analysis had been valid.

Table 2

Characteristics of Five Sites and their Analysis by Thomas and Collier, 2002

						Characteristics of a Scientific Study			
District or School	Location	Program(s)	Students	Relevance[1]	Assignment Process	Control Group[2]	Statistical Analysis	Variables in Statistical Analysis	Validity of Statistical Analysis
Madawaska District and Maine School Administrative District #24	Rural Maine	French Enrichment	English speakers of French heritage	Minimal	Self-Selection	English speakers in mainstream who did not select the program	ANOVA: Stepwise regression assessing increment in r^2	1) Socio-Economic Status (SES); 2) years in program (0-mainstream)	NONE – no information on direction and strength of relationship (b coefficient) or significance of variables
Houston, Texas Independent School District	Urban Texas	Bilingual Education - Developmental (DBE): 2-Way, Transitional (TBE)	LEP; native English speakers	High	DBE, 2-Way: Self-Selection; TBE: Assignment	LEP and native English students in Alternative Programs	NONE	NONE	NONE – no statistical analysis
Grant Community School	Inner-city, Salem, Oregon	2-Way Bilingual Education	LEP; native English speakers	Low	Self-Selection	NONE	Stepwise regression assessing increment in r^2	1) SES; 2) years in program	NONE – no control group; no information on direction and strength of relationship (b coefficient) or significance of variables
Anonymous District	Medium size, Urban, Southeast	ESL	Former LEP	Low (no control group)	District Assignment	NONE (no alternative programs)	Multiple Regression	Nine variables measuring characteristics of students before and at entry to district	NONE – no control group

[1] Relevance to the debate over the best way to educate immigrant limited English proficient students.
[2] Similar students in alternative program(s).

In fact, their analysis of the French bilingual immersion programs used an invalid stepwise regression approach in which the importance of socio-economic status and years in the program were assessed by the increment in r^2. No other data was presented, such as b coefficients, Betas, t values, and significance level so we don't even know the direction of the relationship. It could be that more years in the program lowers achievement.

The third site is Houston, Texas. Different Spanish bilingual programs are examined, but there is no statistical analysis to determine the effect of the different programs on student achievement. Furthermore, throughout this section, Thomas and Collier present simple, descriptive tables of different students in different grades in the same year, and then draw conclusions about "declining" or "increasing" achievement of students over their schooling. To do so when measuring only one point in time is, at best, misleading, at worst, deceptive.

In the Grant Community School in Salem, Oregon their analysis of a Spanish-English two-way immersion program suffers from two fundamental problems. First, there is no control group – similar LEP students in alternative programs. Second, as with the French bilingual immersion study, the only statistical analysis is invalid because Thomas and Collier use the increment in r^2 as the standard for assessing the importance of socio-economic status and years in the program. No other data is presented so the relationship could be negative between years in the program and achievement.

The fifth site is a district in the Southeast that does not want to be identified, a fact which does not invoke confidence in the results and conclusions. Furthermore, there is only one program in this district – an ESL program. Once again, one can only wonder why

Thomas and Collier selected this district. They do not explain. To compensate for the fact that there is no alternative program, they look at many variables that purportedly measure the characteristics of the students before and on entry to the district. These variables are unlikely to be accurate, indeed one is fraudulent – and cannot compensate for a control group.

The analysis that should have been conducted was a multiple regression analysis of LEP students in different programs – mainstream, mainstream with daily ESL pullout, two-way immersion, developmental bilingual education, and transitional bilingual education – controlling for student and school characteristics that also affect achievement. The districts should have been selected either because they had all those programs or because they could be combined to yield a single data file with all these programs, controlling for the district effect.

Instead, Thomas and Collier had the goal of representing different geographic areas of the U.S. That would not have been a problem except for the fact that, with the exception of Houston, they were lacking alternative programs. Thomas and Collier do not appear to know how to do valid research and analysis and unfortunately, they have a large audience in the U.S. that does not understand that they do not understand.

Dr. Christine H. Rossell is Professor of Political Science and former chairman of the Political Science Department at Boston University. She has been conducting research and writing on school desegregation for more than 30 years and on bilingual education for more than 25 years. She has written five books and over 100 articles, book chapters, and technical reports on school desegregation, bilingual education, and other educational issues.

References

Bracey, Gerald. Those Misleading SAT And NAEP Trends: Simpson's Paradox At Work, 2003. http://www.america-tomorrow.com/bracey/EDDRA/EDDRA30.htm

Collier, V.P. (1987). Age and rate of acquisition of second language for academic purposes. *TESOL Quarterly, 21,* 617-641.

Collier, V.P. (1989). How long? A synthesis of research on academic achievement in second language. *TESOL Quarterly, 23,* 509-531.

Collier, V.P. (1992). A synthesis of studies examining long-term language minority student data on academic achievement. *Bilingual Research Journal, 16*(1-2), 187-212.

Collier, V.P., & Thomas, W.P. (1989). How quickly can immigrants become proficient in school English? *Journal of Educational Issues of Language Minority Students, 5,* 26-38.

Cummins, J. (1981). Age on arrival and immigrant second language learning in Canada: A reassessment. *Applied Linguistics, 1,* 132-149.

Hakuta, K., Butler, Y.G., & Witt, D. (2000). How long does it take English learners to attain proficiency? *University of California Linguistic Minority Research Institute Policy Report 2000-1.* Santa Barbara, CA: University of California-Santa Barbara. http://www.lmri.ucsb.edu/

Howard, Elizabeth R. and Julie Sugarman. (2001) "Two-Way Immersion Programs: Features and Statistics," Center for Applied Linguistics, http://www.cal.org/resources/digest/0101twi.html.

Intuitor.com. Simpson's Paradox — When Big Data Sets Go Bad, in Amazing Applications of Probability and Statistics at www.intuitor.com.

Moore, David S.and McCabe, George P. *Introduction to the Practice of Statistics,* W.H. Freeman, 2002.

Rossell, Christine H."Mystery on the Bilingual Express: a Critique of the Thomas and Collier Study," Read Perspectives, V (2), Fall 1998: 5-32. http://www.bu.edu/polisci/people/faculty/rossell/papers/ThomasCollier.pdf

Rossell, Christine H. "Different Questions, Different Answers: A Critique of the Hakuta, Butler and Witt Report,'How long does it take English learners to attain proficiency?'," READ Perspectives, Volume VII, October 2000: 134-154. http://www.bu.edu/polisci/people/faculty/rossell/papers/RossellHakutaCritique2000.pdf

Rossell, Christine H. and Keith Baker, "Selecting and Exiting Students in Bilingual Education Programs," Journal of Law and Education, 17 (4), Fall 1988, 589-624. http://www.bu.edu/polisci/people/faculty/rossell/papers/JournLaw&Ed1988.PDF

Thomas, Wayne P. and Virginia Collier. 1995. "Research Summary of Study in Progress: Results as of September, 1995, Language Minority Student Achievement and Program Effectiveness." http://lexingtoninstitute.org/docs/Thomas++Collier+Summary+1995.pdf

Thomas, Wayne P. and Virginia Collier. 1997. "School Effectiveness for Language Minority Students." NCBE Resource Collection Series, No. 9, December 1997. George Mason University.
http://lexingtoninstitute.org/docs/Thomas++Collier1997%20document.pdf

Thomas, Wayne P. and Virginia Collier. 2002. "A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement." Center for Research on Education, Diversity and Excellence.

Lexington Institute

1600 Wilson Boulevard, Suite 900
Arlington, VA 22209
Tel: 703.522.5828 Fax: 703.522.5837
www.lexingtoninstitute.org
mail@lexingtoninstitute.org