An Analysis of the Research Methodology of the Ramírez Study

Wayne P. Thomas
Graduate School of Education
George Mason University

Abstract

This review of the Ramírez study provides an analysis of the political, educational, and technical factors that strongly influenced the study's research design. The study's data collection and analyses represent a procedural compromise among competing interests of the stakeholders, the requirements of defensible research practice, and the limitations imposed by finite resources and existing U.S. language-minority programs. Analyses of the strengths and weaknesses of the study's research methodology are provided, along with their implications for decision-making in language-minority education. The last section provides a summary of defensible conclusions of the Ramírez study that have not yet been emphasized, in order to clarify incorrect interpretations of the data and to maximize the useful information to be gleaned from this important research effort.

Introduction

The *Longitudinal Study of Structured English Immersion Strategy, Early-Exit, and Late-Exit Transitional Bilingual Education Programs for Language-minority Children*, hereinafter referred to as the Ramírez study (Ramírez, Yuen, Ramey & Pasta 1991, Vol. I; Ramírez, Pasta, Yuen, Ramey & Billings 1991, Vol. II), is a federally financed longitudinal attempt to compare three types of instructional programs for language-minority (LM) students that differ from each other primarily in the degree that teachers provide support in the student's native language as the student learns English and content area material. The Ramírez researchers initiated this eight-year study in 1983, collected data through 1988, and completed the study in early 1991. This study compares the most common bilingual program type, early-exit, with two alternatives, structured-immersion and late-exit. The two alternative programs were little used in 1983 but federal officials deemed them worthy of a formal investigation to ascertain their potential for helping LM

students to learn English and function in the instructional mainstream during their school years.

Several major categories of factors appear to have influenced the research design, the study's plan for data collection and data analysis. These include political, educational, and technical factors, as well as criticisms of past studies. Each type of influence is presented and discussed separately in this paper. Each of these categories of influences led to specific features of the research design or approaches to data analysis in the Ramírez study. First, the study's research design had to allow for the comparison of LM student achievement to that of native-speakers in order to address the educational question, "Will LEP (Limited English Proficient) students ever catch up to native-speakers in the sustained effects of schooling, such as English proficiency and overall long-term achievement levels, to function at least adequately in the world of the early 21st century?" Second, in order to address political and educational concerns, the study's design provided for a direct comparison of the three instructional alternatives that would allow for a defensible judgment as to which of the three alternatives might afford the best opportunity for typical LEP students to match eventually the achievement of typical native speakers. Third, it was necessary that the research design allow types of data collection and types of data analyses that would withstand external scrutiny to a sufficient degree to be credible to the various stakeholder groups with different primary interests in the study, some of whom opposed each other. In other words, the study's methods and features had to make sense at least to both laypersons and specialists before any consensus on policy decisions could be reached. Finally, it was necessary that the research methodology improve on past studies' procedures and, at the same time, adhere rigorously to measurement and analysis guidelines that would pass the scrutiny of the research community.

In the final analysis, the data collection and data analyses carried out in the study's research design represent a procedural compromise among competing interests of the stakeholders, the requirements of defensible research practice, the limitations imposed by finite resources, and the limitations in the types and numbers of LM student programs that existed in U.S. school systems in 1983. Like most compromises, the research methodology satisfies no interested party completely. The researchers themselves express regret in several instances that additional promising analyses or

approaches could not be followed up because of time and resource limitations.

Since the defensible conclusions of any study are constrained by its research design features, it is no surprise that the conclusions of the Ramírez study do not answer definitively all questions of interest to all parties. However, the study does represent a substantial advance in the quality of its analyses, compared to prior studies, and does provide some defensible conclusions that can be used to make policy decisions. Of course, emphasizing one defensible conclusion over another in making the informed judgments necessary for policy decisions can lead to substantial differences in what various stakeholder groups consider important about the study. This paper will draw attention to some defensible conclusions that have received attention and will point to other defensible conclusions that have not been emphasized as yet in an attempt to maximize the amount of useful information that can be gleaned from this important research effort.

Influences of prior studies on the research design

The research design of the Ramírez study was constructed with foreknowledge of the criticisms visited upon large-scale evaluation research conducted in the 1970s. One of the lessons that the Ramírez study researchers learned from past studies was to provide for full specification and description of the instructional interventions under study. The American Institute for Research (AIR) study of bilingual programs (Danoff 1978) was criticized for including a wide variety of types of bilingual programs without adequate differentiation among them and without verifying that these programs really were faithful to a predetermined definition of the characteristics of such programs (Gray 1977; O'Malley 1978; Swain 1979). This means that the AIR study failed to allow for considerable variation in the conditions and instructional strategies that are used in bilingual programs.

In the Ramírez study, the researchers excluded from the study those programs that did not exhibit fidelity with the defined characteristics of structured-immersion, early-exit, and late-exit programs. The Ramírez researchers gave considerable attention to describing the programs nominally, referring to the program s characteristics and planned interventions. They also described the programs operationally, by collecting process data on instructional strategies, procedures, and methods actually used in the classrooms.

In doing so, the researchers pointed out instances of programs that were nominally in one category (e.g., late-exit) but operationally in another (e.g., early-exit). These descriptions allowed for more precise comparisons of the outcomes of programs that adhered to a certain set of characteristics. In addition, the enormous amount of process data collected in the Ramírez study allowed the researchers to compare the programs in terms of the degree of their use of native language, types of student-teacher interactions, degree of teacher training, and many other instructional process variables, so that judgments could be made as to the degree that each program was living up to its full theoretical potential to produce the desired instructional outcomes in LM students. These resource-intensive data collection and analysis activities carried out the purposes of an implementation evaluation and a process evaluation, as described in the classic Discrepancy Evaluation Model (Provus 1971).

The AIR study was also criticized for failing to address the possibility of differential selection (i.e., initial differences between compared groups that may affect comparisons of final group outcomes) in its comparison of treatment and comparison groups (Gray 1977; O'Malley 1978; Swain 1979). The Ramírez study addressed this problem by means of a thorough investigation of the ways that students in program types were different from each other at pretest time and by use of analysis of covariance, with the effects of pretest and other covariates that were determined to be different for the three groups statistically removed from the posttest scores. This resulted in a set of adjusted posttest scores, each adjusted upward or downward when the predicted effects of each covariate had been removed statistically. These adjusted scores yielded adjusted means, each higher or lower than the unadjusted means.

Another serious criticism of the AIR study was the unrealistically short five-month period that was measured with pre-post test scores to determine the effects of program treatment (Gray 1977; O'Malley 1978). During the early 1970s, Canadian bilingual education researchers amassed substantial research evidence demonstrating the importance of long-term measures (at least 4-5 years) to reach an adequate understanding of students' performance in a second language (L2). The Ramírez study collected student achievement data over a four-year period, a considerable improvement over previous studies. However, even with this lengthened period of data collection, the Ramírez researchers found that results possibly attributable to program differences were not

apparent until the fourth year, and more data over succeeding years was needed to answer the program effectiveness questions of the study more definitively.

The analysis procedures of other past studies have also been criticized for use of test scores not on an interval scale (e.g.,grade-equivalent scores or percentile scores), lack of use of measures to equate experimental and comparison groups by language proficiency upon entry into the program, loss of subjects as the program continues over time, and invalid and unreliable testing instruments (McLaughlin 1985; Willig 1985). All of these potential problems are addressed by the research design of the Ramírez study.

Political influences on the research design

In the U.S., there has been controversy as to the best way to provide schooling for LM students for many years. In the past twenty years, much of this controversy has been fueled by changes resulting from court orders, state legislation, and varying emphases in federal funding for educational programs for LM students. The decision to conduct the Ramírez study was one federal government response to this controversy. Federal officials were interested in criteria of instructional efficiency and relative cost, especially since federal funding of early-exit programs was being questioned by the Reagan administration.

However, a variety of other stakeholder groups had somewhat different primary criteria for determining the relative worth of these instructional alternatives. Educators and parents were interested in finding a program that offered sufficient instructional effectiveness to allow LM students to catch up eventually to the native-speaking students with whom LM students must compete for jobs, college admission, and other long-term life goals. Linguists and academicians were interested in testing the theory that supports each of these instructional alternatives in the "real world" and in further developing these theories in order to improve the effectiveness of the delivered instructional "product". Proponents of immigrant and ethnic group interests followed the study with respect to equity issues. They were especially interested in its implications for addressing their concerns about negative effects of typical classroom instruction on their students' cultural values, self-esteem, and long-term equal opportunities with those of native speakers.

There was strong interest by federal officials in a direct comparison between structured-immersion programs and the most

commonly funded type of bilingual program, early-exit (more commonly referred to as transitional bilingual education). As analysis of possible sites for the study was conducted, it was recognized that there was extensive variation among programs nominally labeled as transitional bilingual education in the number of years of first language (L1) support. The decision was made to distinguish between early-exit and late-exit programs and to measure the effects of late-exit (with L1 support for K-6) on LM students as a third distinctive program model. Two other types of programs for LM students were not included in the study design, for unknown reasons--English as a Second Language (ESL) (providing no L1 support, with students of many language backgrounds in each class), and two-way bilingual education (in which language majority students are included in the bilingual class and both language groups study academically through their two languages).

The federal pressure to include structured-immersion as a program model severely limited the study design. In 1983, there were only a few kindergarten structured-immersion programs in the U.S.; thus new sites just beginning the program at kindergarten level provided the only available research locations and limited the study design by effectively restricting the focus to grade levels K-3 for the four years of analysis. Since it is mainly in the upper elementary grades that the curriculum becomes academically more demanding, a much more meaningful analysis of program effectiveness would have been an examination of Grades 3-6 across all three programs, including students who had received the program treatment and continued in the mainstream. The serious implications of this problem in the study are explained in the following section.

In addition, the decision to choose some sites that implemented *both* structured-immersion and early-exit programs also severely limited the external validity of the study design, as the sites did not represent a random or representative sample of the universe of early-exit programs, the most widely implemented bilingual program model. Thus, any conclusions from the comparison of structured-immersion and early-exit programs in the few schools in which the rare structured-immersion programs were found may adequately represent the small national population of structured-immersion programs but probably do not describe the national population of early-exit programs. The researchers recognize this when they state that conclusions from the structured-immersion vs. early-exit comparisons may be generalized to “programs serving Spanish-

speaking language-minority students ... that exhibit the same characteristics as the study programs selected." In addition, they add, early-exit sites in the study "are not representative of all early-exit programs" (Ramírez et al 1991, Vol. II, p. 92).

Educational influences on the research design

An educational factor, the plight of the LM student, is central to the justification for the Ramírez study and for some of the features of its research design. The difficulties faced by students who receive schooling in a language other than their native-language can be substantial. They must succeed in acquiring a new language well enough to conduct basic communication, avoid falling behind in their content area studies (e.g., math, science, social studies) while learning academic English, and then continue to improve the quality of their English enough to be successful in more cognitively demanding tasks of later schooling. In order to have equal opportunity with native speakers, the typical LM student's performance must eventually match the performance of the native speaker in both cognitively easy and demanding tasks before graduation. This implies a direct comparison between language-minority performance and native speaker performance in school-based achievement, a major and valuable feature of the Ramírez study's trajectory analysis of matched percentiles (TAMP) analyses.

In the Ramírez study, this need for the data analyses to compare LEP students' long-term achievement to that of native-speakers becomes even more obvious when one examines the LEP student's achievement experience more closely. The student whose primary language is not English faces a daunting task in attempting to master English at a level that will allow him or her to compete successfully with native speakers of English in advanced academic work in hi h school and in higher education. While day-to-day communication needs can be met with relatively little English instruction, success in school or job in more complex writing, communication, and conceptualization tasks requires many years, even for the typical native speaker. As the LEP student is mastering English, native-English-speakers are also continuing to refine their understanding of the language and are doing so efficiently, because they are taught in the language they know best. Also, there is the potential that LEP students will fall behind the native-speakers in content area instruction because they must devote a substantial fraction of the available instructional time to learning English while the native-

speakers have the luxury of spending this instructional time on advancing their learning in math, science, and other instructional content areas. This combination of lifelong experience with English, plus increased learning efficiency because of instruction in their native language, plus the lack of a need to devote extra instructional time to learning English, gives the native-English-speaking student a large and continuing advantage relative to the LEP student. Thus, in the long term, the LEP student must acquire English even faster and more efficiently than the native speaker, while not falling behind in content area instruction, in order to catch up eventually in proficiency and in overall achievement. In other words, the native speakers start out ahead, are instructed in their native language from the beginning, may learn faster and more completely while the LEP students are acquiring English, and are a ''moving target'' in that they do not slow down in their continuing acquisition of English skills and instructional content so that LEP students can catch up easily. Even in years when the LEP students advance as much as the advantaged native speakers do, they only maintain the existing achievement gap but do not close it as the native speakers continue to advance. Only when the LEP students consistently out-gain the native speakers over several years can they ever catch up.

The impact of these matters on the research design of the Ramírez study apparently was not fully considered by the federal officials who wrote the study's specifications. To see how this is so, we must imagine a hypothetical LEP student, who begins typically in the bottom one-fifth of the national distribution of English achievement. This level of achievement corresponds approximately to the 20th-30th normal curve equivalent (NCE) or the 8th to 17th percentile in a normal distribution. Although the test administered in English may initially underestimate LEP student achievement during the early years of L2 instruction, at some time the typical LEP student's achievement is at the 30th NCE (17th percentile). The typical native-speaking student of similar age and development scores at the 50th NCE (50th percentile) of the national norm group by definition. Thus, in order for the typical LEP student ever to catch up to his or her native-speaking counterpart and eventually close the 20 NCE achievement gap, the LEP student at minimum must match the native-speaker's gains, just to keep the native-speaker from widening the achievement gap even more. This, by itself, is a difficult task. However, in order to close the

achievement gap over time, the LEP student must exceed the native-speaker's gains each year for a number of years, and then continue to do so each year as the instructional material increases in difficulty with each passing grade. For example, the LEP student could catch up by making "normal" gains (i.e., a gain equivalent to that of the typical native-speaker) plus out-gaining the native-speaker by 4 NCEs (a gain equivalent to approximately one-fifth of a standard deviation) for each of 5 consecutive years and then maintain gains equivalent to those of native-speakers until graduation. This pattern of sustained gains over and above those of the constantly advancing native speakers is what is necessary for the typical LEP student to finish 12th grade at the achievement levels of the typical native-speaker.

LEP annual program gains of 5 NCEs (about one-fourth of a standard deviation), when measured in a spring-to-spring testing program that utilizes sound evaluation practices, are considered evidence of moderate success in an instructional program. Thus, even assuming an optimistic student growth rate, we can realize that it is to be expected that typical LEP students should require at least 4-5 years, and probably more, to close an achievement gap of 20 or more NCEs. Therefore, any study that compares the relative efficacy of LEP instructional approaches should follow the progress of those students over a 5-6 year period at least in order to document that the gap has been closed and that the LEP students do not fall behind again after closing the gap. Whether the students are in a structured-immersion, early-exit, late-exit, or other program for LEP students or whether they are in the mainstream, their progress can be expected to be long-term rather than short-term.

Because of the shortsightedness of federal officials, the Ramírez study methodology failed to provide adequately for this expectation of sustained growth, examining structured-immersion and early-exit programs only in Grades K-3. For late-exit students, separate student cohorts were followed from Grades K-3 and from Grades 3-6. The study's decision-making value would have been enhanced considerably by following both structured-immersion and early-exit groups from Grades 3-6 as well, especially as they left their LEP instructional programs and entered the mainstream. The decision not to collect data from a Grade 3-6 cohort for the structured-immersion programs might be justified by the small number of participating students. However, post-third grade data for many early-exit programs was available but not collected.

In summary, the Ramírez study appropriately focused its TAMP analyses on the descriptive statistical comparison of LM student performance in each instructional program to the performance of the national norm group, representing mostly native speakers. However, it failed to provide adequately for the long-term data necessary to compare the school-linked and district-linked structured-immersion and early-exit programs beyond third grade in its hierarchical linear model (HLM) and TAMP analyses. Although the specifications for these two programs called for students' exit to the mainstream before third grade, the federal officials who wrote the research contract specifications might have surmised that a longer term examination was appropriate, in order to ascertain that LM students' gains were sustained after they entered the mainstream.

Technical influences on the research design

Three types of technical factors influenced the Ramírez study's methodology for data collection and data analysis. First, its research design provided for some technical criticisms from past studies such as the AIR study (Danoff 1978). Second, the requirements of sound research practice and the opportunities afforded by new developments in the analysis of multi-level longitudinal data affected the researchers' choices in methods of data collection and analysis in a number of ways. Third, some technical issues that affected the later analyses of program impact (Phase II of the analyses) emerged only after the researchers completed initial program descriptive analyses (Phase I).

The initial descriptive data analyses (Phase I) of the Ramírez study addressed areas of technical concern in the AIR study (Danoff 1978). The first research question, "to what extent does each of the instructional programs in this study reflect its respective instructional model?," dealt with the potential problem of inadequate differentiation among the program treatments. These analyses focused great attention on fully describing the three instructional strategies in terms of how they were planned and actually implemented in the classroom. The preliminary descriptive analyses offered a wealth of information on how these strategies actually differed from each other, especially in terms of proportion of English used in the classrooms of each program.

After a thorough comparison of the three programs with respect to possible differences among them, other than the defined

characteristics of the programs, analysis of covariance techniques were also used to statistically adjust final posttest scores (e.g. in Grade 1) for the effects of pretest scores (e.g., in kindergarten) plus a variety of covariates, including parents' education, socioeconomic level of parents, number of books in the home, and other variables gleaned from parent interviews and school-based information sources. In order to avoid removing posttest variance that might be attributable to the program, thus inappropriately adjusting the program effect, only parent-related and school variables not related to the choice of instructional program were used as covariates, in addition to student pretest score.

The research plan for data collection and analysis was also influenced by new technical developments in the study of longitudinal, multi-wave individual student data (Willet 1988; Williamson, Applebaum, & Epanchin 1991) collected from multiple levels that were hierarchically organized into students, classes, schools, and districts (Raudenbush 1988). Hierarchical linear models (HLM) of individual growth curves as expressed in scaled standard scores [called expanded scale scores in the Comprehensive Tests of Basic Skills (CTBS) used in the study] were utilized to compare the three programs and to control for school and district effects where possible, in order to estimate the true program effect associated with each type of program. In Grades K-i, a traditional analysis of covariance was conducted, while in the analyses involving Grades 1-3 and 3-6, the HLM analyses used individual student growth curves, since there were multiple measurements of student achievement available over time.

An additional technical influence on the study's methodology was caused by longitudinal data collection over a six-year period (the specified length of instruction for late-exit programs), even though the study was funded for only four years of data collection. Therefore, it was necessary to use groups of students who began in each instructional program at the same time but in different grades in order to cover seven grades (K-6) with only four years of data. Because the study was intended to examine closely the effects of structured-immersion and early-exit on students who would be quickly mainstreamed, the researchers gathered data on a cohort of students who began in kindergarten (K-Grade 1 cohort) and on another cohort that began in first grade (Grade 1-3 cohort). Since it was planned that the late-exit students would remain in the program beyond grade four, cohorts beginning in kindergarten (Grade K-3

cohort) and in third grade (Grade 3-6 cohort) were used for late-exit programs.

A final technical influence on the study emerged only after preliminary analyses of the programs' characteristics had been completed. As rich descriptive data was being collected for each of the three instructional program types, the researchers discovered that the three programs actually exhibited eight different "patterns", as defined by the degree to which they varied in their use of English in instruction. Since the proportion of English used in instruction was the primary means of distinction among the three investigated programs, the discovery of distinctly different patterns of instructional English use among the early-exit and late-exit programs led to the opportunity to analyze this variable's range of variation with respect to student achievement growth in each variation. However, after noting that one ostensibly early-exit site was virtually indistinguishable from a structured-immersion program, and that one late-exit site more closely resembled an early-exit program, the researchers failed to examine these instructional patterns further, citing resource limitations. Consequently, the comparisons between programs were based on the programs nominal definitions, rather than their operational definitions, resulting in reduced distinctions among the actual program treatments.

In summary, a number of political, educational, and technical influences on the study's research methodology may be observed. Each of these influences led to choices in data collection and data analysis, that taken collectively, defined the specific features of the research design and methodology of the study. In the next section, some of these features will be examined retrospectively, and possible improvements and alternatives to the choices actually made will be suggested.

Features of the research design and data analysis with possible improvements

The overall design of the Ramírez study was quasi-experimental. Data was collected on LM students while the students' instruction was occurring between 1984 and 1988. The primary comparison variable of interest was type of program, whose levels included the three pre-determined instructional program types to be compared. The study's database combined quantitative outcome measures with a variety of mostly qualitatively-collected and descriptively-analyzed

data on the characteristics of the students, their home background, and their instructional context.

Although much of the descriptive data was collected by means of classroom observations, parent interviews, teacher and administrator interviews, and other information sources characteristic of qualitative studies, most of the questions posed appear to be primarily pre-determined rather than emergent. The researchers utilized qualitative sources of information as a ‘‘data screen’’ to discover issues that were worth closer investigation, typically using quantitative means of data analysis. The primary intent of the initial descriptive phase of the analysis was to provide a case-study presentation on the students' instructional context, on the degree to which the three treatments were implemented in adherence to their specifications, and on the comparability of the three programs in terms of their strategies and site characteristics. In other words, the qualitative information collected was utilized more to document group comparability than to achieve an anthropological understanding of the phenomena being observed.

Unit of analysis. The study chose the individual student as the appropriate unit of analysis because students do not stay in the same classes over the years and because many variables that successfully predict student achievement are at the individual student level (e.g., socioeconomic status of the student). Since student data aggregated to the single levels of class, school, or district can cause ecological effects that can be confused with true program effects, most of the analyses were performed on longitudinal test scores of individual students who were analyzed as being “nested” within classes, schools, and districts. In this way, the researchers could investigate the important effects that might operate simultaneously at each level of schooling, making it possible to avoid the problems of aggregation bias and instability in the estimates of parameters for variables such as instructional program effect (Raudenbush 1988).

The sample. Since the students participating in the study came from programs in existing schools and school districts, there could be no random assignment of students to treatments. Because structured-immersion programs were uncommon in 1983, there was no reason to sample randomly from the population of nine U.S. school districts that had structured-immersion programs (as well as early-exit programs) that could be found to adhere adequately to the

program definition and specifications. Five of the nine eligible structured-immersion-and-early-exit sites agreed to participate in the study, four in the first year of data collection. Thus, the initial sample for comparing structured-immersion and early-exit strategies consisted of four self-selected schools, three in California (two in one district) and one in Texas. In these groups, 139 structured-immersion students and 67 early-exit students were followed longitudinally from Grades 1-3.

In order to supplement the structured-immersion vs. early-exit comparison, the researchers also selected districts in which either structured-immersion or early-exit programs existed in a given school. They selected additional one-program structured-immersion and early-exit school sites from the same California and Texas school systems that had provided the two-program sample. In addition, more one-program sites were selected from one school district in Texas and two in New Jersey. In the one-program sites, a total of 194 structured-immersion students (53% from New Jersey, 17% from California, and 30% from Texas) in 16 schools and four districts were compared to 252 early-exit students (26% from New Jersey, 38% from Texas, and 36% from California) in 13 schools and five districts as a part of the longitudinal analyses of student performance from Grades 1-3.

As in the case of structured-immersion sites, there were few sites nationwide that met the study's criteria for a developmental primary language (late-exit) transitional bilingual education program; none of these had either structured-immersion or early-exit programs. Therefore, there could be no site linkage between late-exit and the other program alternatives. Of the five districts that were located nationwide, three districts agreed to participate in the study, contributing a total of 170 students to the Grade 1-3 longitudinal analyses. Of these students, 20% were from a Florida school, 50% were from seven schools in a New York district, and 29% were from six schools in a California district. An additional 154 students (14% from Florida, 64% from New York, and 23% from California) were followed from Grades 3-6 in the late-exit study.

External validity. Thus, the study's sample for comparing structured-immersion and early-exit programs in the same schools severely constrains the generalizability of the results from these comparisons because of the small number of eligible school districts

analyzed. While the five structured-immersion schools may (or may not) adequately represent the nine nationally eligible structured-immersion program sites, it is unlikely that the paired early-exit programs at these sites are representative of the large number of early-exit sites nationwide. These same-school comparisons do allow for the separation of district and school effects from program effect with a high degree of internal validity, but the degree to which these conclusions apply to comparisons of structured-immersion and early-exit programs at-large is questionable.

In the one-program-in-a-school comparison of structured-immersion and early-exit programs, it was possible to separate district differences but not school differences from the program effects. Moreover, while the structured-immersion students in these schools might have adequately represented the small number of structured-immersion programs nationwide, it is unlikely that the early-exit schools were representative of early-exit schools nationwide.

In summary, this study analyzes data from a majority of the U.S. schools that had both structured-immersion and late-exit programs that fit the study's program criteria in 1983. Thus, the researchers claim that these samples adequately represent the variety of such programs in the first year of the study. However, for early-exit programs, only those that existed in conjunction with structured-immersion programs in the same school or in the same district were sampled. Thus, we have no good knowledge of the range of possible outcomes to be expected nationally from early-exit programs, the most commonly funded type. Our conclusions are limited by what can be said about schools or districts that had both structured-immersion and early-exit programs.

Since no schools with both early-exit and late-exit or with both structured-immersion and late-exit could be found, it was necessary to sample separately those schools and districts that did have late-exit programs. Thus, the researchers correctly assert that the sample of late-exit programs is not directly comparable to the sample of structured-immersion and early-exit programs that are linked by common schools or districts. Because the late-exit programs shared neither districts nor schools with the other two program types, any differences in student achievement between the two samplings would always contain district and school effects, in addition to program effects. These effects would be confounded and virtually impossible to separate out in an inferential analysis.

The researchers argued that the only analysis that could separate out program effects was a comparison between the structured-immersion and early-exit programs that occurred in the same schools, and thus districts. In addition, they stated that comparisons among the different districts that conducted late-exit programs could control for district effects, but differences between schools remained confounded with program effects. While these assertions were well founded, they sidestepped the fact that a study conducted for policy-making purposes needed to exhibit some generalizability of findings as well as to seek internally valid conclusions in comparing programs.

A more fundamental concern arises as to why scarce federal research dollars were used to investigate the apparently rarely-used structured-immersion strategy in the first place. The most commonly used LM instructional strategies in the early 1980s were ESL and transitional bilingual education (early-exit) programs. For policy making purposes, surely the most commonly used or commonly funded strategies should have been most thoroughly investigated. In addition, one might justify examining the effects of late-exit programs because, of the U.S. alternatives, they most closely resemble the Canadian immersion programs whose success with language-majority students is well documented (Ramírez et al 1991, Vol. I, p. 26).

Thus, a random sample of ESL, early-exit, and perhaps late-exit programs would have better served the decision-making needs of this study's stakeholders. Even a purposive sample of typical and exemplary programs in each of these categories might have furnished more useful information concerning the relative effectiveness of these approaches. At least, the sampling plan of the study should have provided for a thorough investigation of the full range of effectiveness of the most commonly funded strategy, early-exit programs. The Ramírez study provides little information as to what should be expected in the typical effects of early-exit programs nationwide.

Thus, on a district-level basis, the sampling plan allows comparisons between early-exit, late-exit, and structured-immersion programs, with school and district effects intermingled with program effects. Also, on a school-level basis, we can draw inferences about the relative effectiveness of structured-immersion and early-exit programs, controlling for school and district effects,

in those few schools that contain the rare structured-immersion programs and also have early-exit programs.

Analyses conducted. Based on theory, the use of HLM analyses was an appropriate choice representing a substantial improvement over the methods of previous studies in the analysis of student growth. However, in practice, the relatively small numbers of students for whom data was available over a three or four year period limited the applicability of these analyses. The hypothesis tests and estimates for HLM models rely on large sample properties of maximum likelihood estimates and little is known of the small sample behavior of the estimates (Raudenbush 1988). In addition, it is not clear whether the researchers tested the parametric assumptions that these techniques do require. However, Raudenbush and Bryk (1988) point out that HLM analyses avoid the problems posed for conventional statistical analysis by heterogeneity of regression. In fact, HLM techniques seek out situations in which student background variables might vary across class or school groups in order to seek explanations for why separate regression lines for each class or school might be related to instructional effectiveness or other class/school characteristics.

The researchers did address these points indirectly by running many analyses whose purpose was to indicate the robustness of the conclusions and their sensitivity to perturbing effects. The researchers varied analytic models and variables to test the sensitivity of the conclusions to the effects of small changes in the ways that the analyses were conducted. Also, they allowed for judgments regarding the degree to which the conclusions were "driven" by the use of certain variables, models, or covariates. The reasons for any substantial changes in results then could be further investigated and assessed in an exploratory manner. Although the overall interpretations of the results from the analyses appeared to be resistant to instabilities caused by varying the covariates, the subjects, or the analytic method, not enough is yet known about the behavior of HLM models with small samples to justify complete confidence in the conclusions. By all available information, the analyses do appear useful and tentatively correct. Certainly, the techniques offer great promise in the appropriate assessment of student achievement growth.

In the Grade K-i analyses, the researchers used an analysis of covariance to adjust the spring Grade 1 CTBS scores for the effects

of such variables as pretest, school, student absences, preschool attendance, average educational level of parents, and number of books in the home. All of these factors represented possible influences on the test scores that the researchers wanted to remove before assessing differences among programs. However, it is not clear whether removing the effects of these covariates may have underadjusted some program means and overadjusted others. The researchers provide little information as to how they tested the ANCOVA assumptions.

In apparent response to the severe limitations placed on policy decisions regarding program comparability by the sampling plan, the Ramírez researchers chose to supplement the HLM analyses with descriptive analyses, using TAMP as described in Braun (1988). These analyses allowed a more direct comparison of the three instructional alternatives than was afforded by the inferential analyses performed on highly restricted samples. In this approach, the full range of student performance in the structured-immersion, early-exit, and late-exit programs was compared, not to each other, but to a common standard, the performance of the mostly native-speaking national norm group. Because of the limitations imposed by the research design, the HLM analyses were not able to compare the effects of the three program types directly, by separating all effects of schools and districts from the effects of programs on student performance. Thus, some method for addressing the three-program comparison was required. Essentially, the TAMP analyses represent a way to compare descriptively the three programs to a common frame of reference, the performance of the 1978 national norm group of the CTBS. The use of the 1978 version of the CTBS was required by federal officials for comparability with other federal studies. It was assumed that any differences in norms (and overestimation of performance) that resulted from using the older norms should affect all three groups equally.

The TAMP charts displayed a plot of the values of two variables, the norm group's performance on the pretest (e.g., the Grade 1 CTBS test in a given subject area) and the norm group's performance on the posttest (e.g., the corresponding Grade 3 CTBS test). However, this was a special form of pre-post plot in which each pretest scaled standard score (referred to as expanded scale scores by the CTBS) was matched with the posttest scaled standard score that corresponded to the pretest score's percentile. In other words, the first percentile scale score for the pretest was plotted

versus the first percentile scale score on the posttest, and this process was continued through all 99 percentiles. Since the expanded scale scores were theoretically on an interval scale and were comparable across the tests for a given subject area, the "equipercentile" plot resulted in a line that defined the performance of the norm group during the interval between the tests. The pre-post scores of each instructional program or group were then plotted on the same plot, resulting in a descriptive comparison of each group's performance to that of the norm group.

Thus, the performance of low-scoring students in each program could be compared to low-scorers in the norm group, mid-range scorers to similar performers in the norm group, and high scorers to their counterparts in the norm group. The result of these plots was a descriptive comparison of the unadjusted scores of structured-immersion, early-exit, and late-exit programs, not to each other, but to the performance of the norm group. This indirectly achieved a major goal of the Ramírez study, to compare the performance of the students in the three programs to each other in that each group could now be compared to a common standard. This standard, the native speaking norm group's performance, was a goal toward which the language-minority students in each program were working.

One possible objection to the TAMP analyses involves their comparison of longitudinal student data to a series of cross-sectional sets of norm group data. The student achievement scores of the participating students represented true longitudinal data, collected on the same students over a period of up to four years. However, the norm group's performance was defined by a series of measurements of the performance of different students in different grades at one point in time, a cross-sectional approach. Since typically there are no longitudinal norms for standardized tests, this comparison with cross-sectional norms was necessary.

It should be noted that a very reasonable criterion for LM student success is the attainment of a match between the performance of norm group students with the performance of LM students in the same grade or age group. The cross-sectional norm group for that year is not only the sole available group for this purpose, it is an appropriate group for these comparative purposes precisely because it defines the state of typical student performance across all grades for the comparison year. The use of a series of cross-sectionally derived "checkpoints" as standards is appropriate when these norms are used to represent the performance of all students (mostly native

speakers) in schools, as opposed to using them to interpret inappropriately the scores of individual LM students, especially those who are new to LM instruction and to a test administered in English. After all, LM students will attain parity in achievement when their distribution of test scores (including measures of central tendency and dispersion) becomes indistinguishable from that of the test's norm group, as they attempt initially to close the achievement gap and then to keep it closed as both distributions of students advance in achievement during the schooling years. In other words, a finding of no-significant-difference between the LM students' achievement, as expressed by the subtests' score distributions, means and standard deviations, and the norm group's achievement (assuming up-to-date norming data) at high school graduation would present prima facie evidence of equity and parity in schooling benefits for students who had participated in LM instruction.

The TAMP analyses provided very useful information, but it should be recalled that both the LM student scores and the scores of the norm group students contained effects due to the schools and school systems where they attended school. In other words, the TAMP analyses were performed using unadjusted test scores, and thus retained the differential effects (if any) of districts and schools. This caused an internal validity problem, in that observed differences among programs may include parts that are attributable to school and district differences. In addition, there remain the previously expressed external validity concerns that the students in the Ramírez study (especially those in early-exit programs) may not reflect the nationwide performance of students in these programs.

The TAMP analyses also included information for computing confidence intervals, indicating the degree of uncertainty inherent in the performance of the students in each program. TAMP analysis is a descriptive technique and does not directly test hypotheses. However, it does allow for confidence intervals to be computed that can provide guidelines for judging meaningful differences between the performance of the norm group and the performance of LEP students in each instructional program over time. For some reason, the study's final report provides information regarding the confidence intervals for TAMP charts in generic and abstract terms of fractions of an inch on the charts provided, rather than including the confidence intervals as part of the graphical information provided in each chart. While this may have been done with the intent of reducing visual clutter in the TAMP charts, the lack of clearly

marked confidence intervals is an impediment to the reader's interpretation.

In all, hundreds of descriptive, HLM, and TAMP analyses were performed on the variety of instructional groups, cohorts, and subject areas. For decision-oriented research consumers, the primary analyses of interest are the HLM analyses that compare the structured-immersion strategy with early-exit in Grades 1-3, the separate HLM analyses that examine late-exit student growth in Grades 1-3 and 3-6, the TAMP analyses that compare the three programs to the national norm group in Grades 1-3, and the TAMP comparison among several late-exit sites in Grades 3-6.

Analyses not conducted. Given the study's extensive efforts to collect process and background data on the participating LM students, it is unfortunate that an analysis of the students' first language cognitive academic development was not included among the primary analyses. Since a key theoretical rationale for bilingual instruction is the transfer of academic knowledge from one language to another, measures of academic development in both Spanish and English should have been included in the program effectiveness questions. This study's conclusion, that the late-exit students are the only group that may be able to close the achievement gap with native-speakers, appears to provide support for the idea that L1 academic development transfers to L2 academic development. However, an analysis of students' L1 cognitive-academic growth would have provided a more direct measure of the function of transfer across languages.

An additional analysis that might have been provided was a descriptive presentation of the three groups progress across Grades 1-3 and 3-6 as expressed in units of normal curve equivalents. While the Ramírez researchers correctly conducted all computations using the CTBS expanded scale scores, most readers of the study would probably have understood the study's conclusions more readily had the results been presented in the familiar format of a federally funded ESEA Title VII longitudinal program evaluation. When the informed reader performs these conversions, especially after reassigning several groups that the study identifies as operationally more similar to another program type, the superiority of late-exit programs in reducing the LEP student achievement gap is more apparent. The study does present these same unadjusted scores in the TAMP charts using expanded scale score units; means

and standard deviations in scale scores are also presented for each group in each grade. Expressed in grade-appropriate NCEs, these achievement means would have furnished more obvious indications that late-exit students tend to close the achievement gap while structured-immersion and early-exit students tend to fall behind the norm group in the long-term.

Defensible conclusions allowed by the research design and analyses

There are several sets of conclusions that may be supported by the Ramírez study. First, the report offers eight implications and conclusions that are intended for use by decision makers. Some of these have received considerable attention in the press and deserve further discussion. There exists also a second category of conclusions that are found in the body of the study's final report that have not received substantial attention as yet. A third category of conclusions consists of those that may be derived from the findings presented in the charts and tables of the study's final report with additional analysis. This section of this article examines a number of conclusions from each of these categories that the research methodology of the Ramírez study can defensibly support. The Ramírez report's question-and- answer format of presenting conclusions works well and is used here for each conclusion analyzed.

Of the three program alternatives, which shows the most promise for helping LM students eventually match native-speaker Performance on standardized tests? When converted to NCEs, the expanded scale score means for each instructional group indicate that only late-exit students may catch up with the typical native-speaker, given a sustained demonstration of the observed gains for five or more consecutive years. The students in structured-immersion and early-exit programs will slowly fall behind the typical native-speakers or, at best, fail to close the achievement gap between them. Both the HLM and the TAMP analyses performed in the Ramírez study suggest the same conclusion. "The HLM analysis showed that the growth curve for immersion strategy and early-exit students was negative, indicating a deceleration in their rate of growth from first grade to third grade. In contrast, the growth curve for late-exit students was positive from

grade one to grade three, suggesting continued growth over this grade span" (Ramírez et al 1991, Vol. II, p. 650).

In fact, late-exit success across sites seems directly proportional to the degree of use of primary language instruction. Based on this author's supplementary analyses, it appears that both structured-immersion and early-exit students can be expected gradually to fall behind the norm group by amounts that fall slightly short of statistical significance over a three-year period. However, these differences, if maintained in cumulative fashion over time, would attain significance after four to five years, or if larger group sample sizes allowed more statistically powerful comparisons.

An additional reason to support the advantages of late-exit programs is provided by the Ramírez report's finding that the late-exit strategy (not structured-immersion) most closely resembles the successful Canadian immersion programs in features and strategies (Ramírez et al 1991, Vol. 1, p. 26). Since the efficacy of the Canadian programs for language majority students has been amply documented, this suggests that programs with the features (not the name) of the Canadian programs may not only be best for language majority students, but also best for language-minority students. In making this conclusion, we look at the operational definition (i.e. what the programs actually do) of the three investigated instructional programs rather than their nominal labels (i.e., what the programs call themselves or claim that they do).

Are structured-immersion or early-exit programs better in providing for LM students' instruction in math. language. and reading? A Ramírez report conclusion that has received some attention is, "...providing a limited-English-proficient student with English-only instruction through grade three, as was done in the structured English immersion strategy program, is as effective as an early-exit program in helping limited-English-proficient students acquire mathematics, English language, and reading skills" (Ramírez et al 1991, Vol. II, p. 664). This finding is defensible only when one restricts the study's domain to schools having both structured-immersion programs and early-exit programs (surely an unusual situation even at present); restricts the study to a period (Grades 1-3) when mostly low-level language skills are taught and all groups gain more than they will with later more cognitively complex instruction; and accepts the reduced power of statistical tests performed on the small groups of students that

remained in the program after four years. An alternate (and equally defensible) form of this conclusion might be expressed as, "In the short term (i.e., three years) only, and only looking at a restricted sample of schools or districts that have both structured-immersion and early-exit programs, there is a three-year trend that early-exit students gain faster than structured-immersion students. With either a slightly larger sample size, or with another year of data, this difference would probably have been statistically significant in the favor of early-exit students."

Both versions of this conclusion are supported by the Ramírez analyses. While neither is incorrect, the first is conservative, is expressed in a restrictive form, and is supportable only when the other analyses in the study are not considered. The second form describes the context of the conclusion, extrapolates from trends observable in the data, and is supported by the author's additional analyses and by other Ramírez analyses. These additional analyses compare the instructional groups as defined by their actual operational patterns of instructional use of English, rather than by their nominal classifications as done by the Ramírez report. Federal officials who wish to fund English-only program alternatives may find the first form of this conclusion attractive. However, educators and others who may be more interested in long-term student academic performance may favor the second form.

Are existing instructional programs operating efficiently from an instructional perspective? Some of the most worthwhile conclusions of the Ramírez study come from the initial descriptive analyses that extensively compared the characteristics of the structured-immersion, early-exit, and late-exit programs. One of the most powerful conclusions is that none of the three instructional approaches is being "all that it can be" instructionally. All are limited to some extent by uninspired teaching or by restricted opportunities for students to produce language and acquire productive skills with English. All offer undemanding instructional environments. There is a low frequency of student-initiated interactions in each of the programs, especially among late-exit students. The study does document that the three programs are equally limited and thus this creates no internal validity problem when the three groups are compared. An obvious implication of this conclusion is that more federal, state, and local resources should be devoted to teacher training and curricular development in order to

address these factors that appear to be limiting the effectiveness of all of the LM instructional programs in the study. Calls for national student tests and teacher tests that do not address these factors should be replaced by calls for in-service training of present teachers as well as preservice training of future teachers that will allow these teachers to develop the specialized teaching skills required for fully effective LM student instruction. If applied to education in general, this policy would amount to a substantial investment in our teaching infrastructure that is much more likely to produce meaningful reform of education than the "quick fixes" now touted so widely.

How much of an effect do the three programs have on student instruction? A conclusion from the HLM analyses is that the effects of each of these programs is smaller than might be expected. This author's supplementary analyses indicate that the larger annual achievement differences among programs are equivalent to approximately one-fifth of a national standard deviation. In most cases, the portion of the student growth than can be attributed to school effects is equal to or larger than student growth attributable to the instructional programs. It is possible that some of the effect attributed to schools may in fact be due more to the particular ways in which a specific program is implemented at that school, resulting in an underestimate of the program effects. In addition, it is possible that programs defined in terms of operational use of L1 in instruction (rather than by nominal definitions) may yield larger program effects. Preliminary re-analyses of the data by this author suggest that program effects may be larger than school effects when the actual operational differences in language use are employed as a key difference between programs.

How can we tell whether these instructional programs are promoting educational equity and parity for language-minority students? The only reasonable (but depressing) criterion for LM-LEP students' eventual instructional success is whether the distribution of low, middle, and high scoring LM-LEP students will ever match the distribution of typical native-speakers, as represented by a test's norm group. Comparison to the norm group makes sense after several years of instruction in English, but less so in the early years of LEP student instruction. However, even then, the norm group can serve as a conservative

and "tough" standard when used to represent long-term group performance goals for LM students.

Does it become more and more difficult with each successive grade for LM-LEP students to avoid falling behind the norm group? The complexity of information and cognitive demand increases with increase in school grade. There are at least two indicators of this. First, the pattern of CTBS expanded scale score means and standard deviations for the norm group is such that the differences between the means become reduced with each passing grade while the standard deviations become larger. In addition, an examination of the items on the CTBS test (and any other nationally-normed standardized test) will indicate that test items tend to sample more cognitively complex skills with more sophisticated usage of English with each passing grade, especially at secondary levels. This observation may explain how LM-LEP students may appear to make quick progress in the early elementary years, even relative to the national norm group, but may quickly fall behind their native-speaking counterparts in the post-elementary school period as their initial acquisition of mostly low-level English skills becomes inadequate to cope with the increasing cognitive demands of the tests, as well as the requirements of more advanced courses that lead to higher education. It is worth noting that instruction carried out at low levels of cognitive demand (especially in the upper-elementary and secondary school years) may riot differentiate between students with "shallow" English skills and those with greater capabilities in conceptualization, written skills, and production of the English language. Thus, the weaknesses of former LEP students in the mainstream may not be apparent until they take a test that is designed to distinguish among levels of performance (e.g., the College Board tests) rather than to document minimum mastery of skills.

However, this depressing picture of the difficulty of "catching up" to native-English speakers on standardized tests does not take into account the potential for transfer of L 1 cognitive-academic development into L2. While the Ramírez study concluded that only the late-exit students might be successful in closing the achievement gap, the study did not include direct measures of the students' L1 academic development. Research evidence from other studies indicates that there is considerable research support for transfer of academic knowledge across languages (Collier 1989; and Collier

article in this volume). Thus, those students with the strongest cognitive/academic development in L1 may have the potential for the highest academic achievement in L2.

Assuming that LM-LEP students do catch up to native-speaker levels of achievement. do they fall behind again? Even if a LM-LEP student catches up to typical native-speakers in the early grades, he or she must continue to make substantial amounts of progress each year to stay caught up. It is worth noting that these students may fare acceptably in classes that are not challenging, that "water down" instructional material, or that consist primarily of low cognitive-demand interactions and activities. Only when LM-LEP students' capabilities for dealing with complex cognitive tasks are compared to those of native speakers may substantial differences between the two become noticeable and even obvious.

Structured-immersion students have an early advantage on tests, since they have been exposed to more English. Thus, a test administered in English tends to be more valid for them early on. The Ramírez study documents this early performance surge relative to the early-exit students by noting that structured-immersion students temporarily move ahead of other LM-LEP students in some subject areas after one or two years of instruction. However, as the early-exit and late-exit students are exposed to more English, and thus the test becomes more valid for them, their initial "lag" disappears. Because the structured-immersion students have sacrificed cognitive development and content in their early emphasis on learning English, their long-term ability to deal with increasingly complex material may be hampered, especially as they enter their years of post-elementary school instruction. This could lead to an elimination of initial gains relative to the norm group and a pattern of sustained losses among students who receive instructional support only in L2, relative to the annual achievement gains of the norm group. The Ramírez study found that the late-exit students, with both L1 and L2 support, were catching up to the norm group even as their academic work became cognitively more complex in the upper elementary grades.

What are the operational characteristics of successful LM instructional programs? The Ramírez study documents that the characteristics of the most successful program included in this

study are: (1) substantial teacher use of the minority language in the early elementary school years followed by approximately 40% (or more) use through fifth grade and 24-26% use in sixth grade; (2) a high degree of teacher proficiency in Spanish; and (3) teachers who have advanced training in meeting the needs of language-minority students. It is worth noting that these variables are ones that can be influenced and changed by local schools, whatever the local choice for LM instructional program.

Are there substantial differences among the three investigated LM instructional programs. other than differences in their use of L1 for instruction? Findings from the extensive classroom observations, surveys, and ratings scales from the descriptive phase of data collection are exhaustively presented in order to specify the ways in which the three programs are different, other than those factors associated with the characteristics of the programs themselves. Relative to the number of variables examined, only a few differences were found between the home, student, teacher, school, and district characteristics associated with each of the three instructional programs. Statistical adjustments were made to adjust for these initial differences and to remove school and district effects, wherever possible, so that program effects on student growth in achievement could be estimated.

Overall, the three programs were found to be comparable with respect to the quality of instruction they provided in amount of instructionally engaged time, complexity and content of student and teacher communication, with a few exceptions involving the late-exit programs. The first of these was that late-exit teachers assigned and corrected homework more often than the teachers of the other two programs. Parents of late-exit and early-exit students worked with their students more than did the parents of structured-immersion students, presumably because they could help them in their native language. While proportionally more late-exit parents were at the lowest income levels, their students were academically the highest achievers in English in the study.

There is at least one case in which significant inter-program differences were expected but not found by the Ramírez study. The Ramírez researchers found that early-exit and structured-immersion teachers tended to retain their students beyond Grade 3 (substantially after their instructional models indicated that they should be

mainstreamed) because of the teachers' perceptions that these students were not ready for the mainstream after 2-3 years of instruction. After four years, only two-thirds had been reclassified as fluent-English-proficient, and only one-fourth of structured-immersion students and one-fifth of early-exit students had been mainstreamed. This "...suggests that LEP students require a minimum of five or more years of special instruction in either structured-immersion strategy or early-exit bilingual programs" (Ramírez et al 1991, Vol. 11, p. 34). Thus, the idea that students should be exited from support services as quickly as possible may be misguided. However, this study provides strong evidence that long-term support in a late-exit program will benefit students the most.

Does the Ramírez study offer suggestions for the conduct of future studies of LM instructional Programs?
More efficient and accurate means of selecting eligible districts are needed in future studies. The Ramírez study addressed very well the initial confusion among the features of programs and among program variations in the multi-stage sample selection process. However, although the study described variations among the early-exit and late-exit programs that blurred the distinctions among these program types, the analyses were carried out using the somewhat undifferentiated nominal descriptions of the programs rather than the operational descriptions. As a result, the distinctions between early-exit and late-exit programs and between early-exit and structured-immersion programs were partially blurred. In these cases, the accuracy of the analyses would have been better served by analyzing the data based on the *operational* definition of programs at each site, given the well-known lack of definitional adherence and confusion of terms in the field. The nominal analyses (comparing groups based on their planned characteristics rather than their actual observed instructional strategies) were probably performed to avoid the political problems of reassigning sites to different programs. However, the nominal program labels are apparently inaccurate in the case of several sites that belong in other categories. This causes confusion in distinguishing differences between the programs and amounts to a "fuzzy" independent variable, something the Ramírez study otherwise tried very hard to avoid.

The researchers could have performed analyses by schools, by school-program combinations, or by the "patterns" of English use

identified in preliminary analyses. However, the numbers of students in each group would have been lower than they are now, resulting in loss of power for the inferential tests and increased uncertainty for the confidence intervals of the descriptive statistical analyses. Although the study collected data from more than 2,000 students, the most meaningful analyses for decision-makers (Grade 1-3 comparisons) were restricted by attrition to a total of several hundred student scores after the initial sample had been followed for four years. Apparently, the Ramírez researchers underestimated the degree of student attrition in a four-year longitudinal study. Future studies should attempt to analyze larger data sets in which larger and more nationally representative groups remain for study after several years of program operation and should include continuing analyses of students' academic progress within the mainstream.

In future studies, the LM-LEP students should be followed for more than three or four years. The Ramírez study shows that most of the structured-immersion and early-exit students were retained in instruction by their teachers even after the students had been reclassified as English-proficient. This implies that neither structured-immersion nor early-exit programs work faster or more completely than late-exit programs that instruct LM-LEP students for six years. The Ramírez study missed an opportunity by failing to follow structured-immersion and early-exit students in a Grade 3-6 cohort. Even if structured-immersion students were in short supply because of the small number of available programs, following early-exit students' achievement descriptively from Grades 3-6 would have provided very useful information about the sustained effects of a commonly funded program that apparently requires more than three years for its effects to be felt.

Summary

The Ramírez study has substantially advanced the methodology of language-minority student instructional program evaluation. While no one analysis is definitive, the report excels at pursuing different lines of inquiry, relating them to the predictions of theory, and presenting a rationale for the analyses to converge on conclusions that are useful for policy-making and informative for educators and parents. Its final conclusions and implications are conservatively worded; however, those who read the full report will find that the researchers did not shrink from exploring prevailing trends and undercurrents embedded in the data that suggest fruitful

courses for further investigation. The study does fail to pursue several promising analyses and lines of inquiry because of resource limitations. However, it also addresses most of the criticisms of past evaluations of Title VII instructional programs. The study's research design and conclusions have been greatly influenced by the political issues that embroil bilingual education. The analysis methodology of the study is sophisticated and well-executed; the disappointments mostly involve smaller-than-expected longitudinal samples and restricted generalizability of conclusions, especially for early-exit programs.

In all, the Ramírez study has provided data that will bear considerable re-analysis and secondary analysis, in the search for a more complete understanding of the key issues in providing a meaningful and appropriate education for language-minority students. This author's attitude toward the Ramírez study is well described by McLaughlin (1985, p. 245) who has suggested, "It can be argued that a great deal can be learned from less than perfect research and less than fully generalizable findings. If one accepts the notion that knowledge in social science grows by accretion, every bit of information contributes to the process. What one must avoid is misinformation, and the more rigorous the research and the more careful the researcher is to deal with the problems that have been discussed here, the greater the contribution to knowledge about the effects of bilingual education." By this criterion, the Ramírez study represents a major, worthwhile contribution to the study of instructional programs for language-minority students.

References

Braun, H.I. 1988. A new approach to avoiding problems of scale in interpreting trends in mental measurement data. Journal of *Educational Measurement, 25*, 171-191.

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

Danoff, M.N. 1978. *Evaluation of the impact of ESEA Title VII Spanish/English bilingual education programs: Overview of study and findings.* Palo Alto, CA: American Institute for Research.

Gray, T.C. 1977. *Challenge to USOE final evaluation of the impact of ESEA Title VII Spanish/English bilingual education programs.* Washington, DC: Center for Applied Linguistics.

McLaughlin, B. 1985. *Second language acquisition in childhood: Vol. 2. School-age children* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

O'Malley, J.M. 1978. Review of the evaluation of the impact of ESEA Title VII Spanish/English bilingual education programs. *Bilingual Resources, 1,* 6-10.

Provus, M.M. 1971. *Discrepancy evaluation.* Berkeley, CA: McCutchan Press.

Ramírez J., S. Yuen, D. Ramey & D. Pasta. 1991. *Final Report: Longitudinal study of structured English immersion strategy, early-exit and late-exit bilingual education programs for language-minority children.* (Vol. I) (Prepared for U.S. Department of Education). San Mateo, CA: Aguirre International. No. 300-87-0156.

Ramírez, J., Pasta, D., Yuen, S., D. Ramey & D. Billings. 1991. *Final report: Longitudinal study of structured English immersion strategy, early-exit and late-exit bilingual education programs for language-minority children.* (Vol. II) (Prepared for U.S. Department of Education). San Mateo, CA: Aguirre International. No. 300-87-0156.

Raudenbush, S.W. 1988. Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics, 13,* 85-116.

Raudenbush, S.W. & A.S.Bryk. 1988. Methodological advances in analyzing the effects of schools and classrooms on student learning. In Rothkopf, E.Z. (Ed.), *Review of Research in Education 15.* Washington, DC: American Educational Research Association.

Swain, M. 1979. Bilingual education: research and its implications. In Yorio, C.A., Perkins, K., Schachter, J. (Eds.) *On*

TESOL '79: The learner in focus (pp. 23-33). Alexandria, VA: Teachers of English to Speakers of Other Languages.

Willett, J.B. 1988. Questions and answers in the measurement of change. In Rothkopf, E.Z. (Ed.), *Review of Research in Education 15.* Washington, DC: American Educational Research Association.

Williamson, G.L., Appelbaum, M., & A. Epanchin. 1991. Longitudinal analyses of academic achievement. *Journal of Educational Measurement, 28,* 61-76.

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