

# **A National Study of School Effectiveness for Language Minority Students' Long-Term Academic Achievement**

## **Research Design**

Our research design is based on a comprehensive data collection effort at each research site, collecting both qualitative and quantitative data that directly address the policy questions of the school district, regarding language minority students and their academic achievement over the long term (4-12 years). We, as well as many other researchers in language minority education, have found that short term research, examining student outcomes for 1-2 years, presents an incomplete and inaccurate picture of language minority students' continuing academic success (Collier, 1992; Cummins, 2000; Lindholm-Leary, 2001; Ramírez, Yuen, Ramey & Pasta, 1991; Thomas & Collier, 1997). Thus the focus of our work is to examine the long-term outcomes in student achievement, following language minority students across as many years of their schooling as is possible within each school district.

We conduct this research at the school district level, collecting data from the central administrative offices, including the offices of testing, bilingual/ESL education, curriculum supervisors, and data processing. We also in each school district collect some school-level data, focusing on visits and interviews with staff and students of individual schools that stand out as promising models of school reform for language minority students, based on their student achievement data. Overall, however, this research could be characterized as providing whole

school district views of policy decision-making that is data-driven regarding designing, implementing, evaluating, and reforming the education of language minority students.

In this process of data collection, the school district staff are collaborative researchers with us. Our initial contact is usually the central administrative assistant superintendent or curriculum supervisor in charge of bilingual/ESL services in the school district. Initial meetings include central administrative staff from the bilingual/ESL and research and evaluation offices of the school district, followed by meetings with the superintendent and associate/assistant superintendents. When all of these parties have agreed to a collaborative research plan, we begin collecting data in that school district. The following overview describes some of the initial processes that are discussed in these first meetings.

## **What We Do with Each School District as Collaborative Researchers: Initial Stages of Study**

Prior to data collection and analysis, we work extensively with our participating school districts to enable them to engage more effectively with us in a multi-year collaborative relationship. In doing so, we introduce our “middle-out” strategy of school reform. Specifically, we:

- Foster a reform climate in each school district by providing professional presentations and consultations for school board members and other policy makers;
- Move the school district towards decision-making based on their own locally-collected data, rather than decision-making based mainly on opinion or political expediency;
- Enable critical staff (mid-level administrative bilingual and ESL staff) to facilitate the change process, through our “middle-out” approach to school reform (rather than top-down or bottom-up approaches);
- Educate and sensitize policy-making staff (central administrators, school board, principals, and resource staff) to pertinent concepts and concerns regarding the education of language minority students;
- Provide an inquiry framework with our general research questions, and encourage school district staff to add meaningful research questions of local interest;
- Introduce and utilize the methodology of program evaluation, based on large-scale studies, with focus on sustained, long-term effects and outcomes (4-12 years), not on the short term (1-2 years). This type of research addresses overall pragmatic concerns of policy

makers, focusing on program outcomes at the school and district levels. Therefore, together we:

- Elicit and clarify local concerns and values;
- Conduct needs assessment;
- Practice formative program improvement and installation prior to summative analysis, to enable full and best implementation practices;
- Acknowledge that most educational effects are small in the short term and practically significant only in the cumulative long term;
- Work with our school district colleagues to decide together on the appropriate data to collect; we advise on data collection methodology and provide technical expertise on instrument development; they collect the data and retain ownership of the data; we analyze the data collected; and we and they collaboratively interpret the results of data findings. As collaborative researchers with us, the school district staff are our “eyes and ears,” and they carry the primary burden of day-to-day data collection;
- Focus our research on large groups of students across program types and across the years, not on small groups studied intensively for a short time. We follow students initially placed in a special program as they continue in the mainstream in later years, to examine their long-term academic achievement across the years;
- Provide pragmatically useful information to policy-makers. At the beginning of our long-term collaboration with each of our participating school districts, we provide the local

policy-makers with information on the long-term outcomes of their local curricular choices, based on data analyses from other school districts. In many cases, this is the first time that policy-makers have had such information to guide their decision-making. We elicit and help clarify local concerns and values and respond to these in our presentations, in our suggested data collection activities, and in our data analyses.

In summary, if the school district is already inclined towards reform, we try to foster that reform climate by providing well-focused questions that local educators and other interested parties should ask of their programs, based on the experiences of other school districts with whom we have worked during the past 10-15 years. We provide a framework for local inquiry about the effectiveness of local schools with our general research questions of interest nationwide, and assist local educators in filling in our national questions with local research questions of interest to them. As data collection stages begin, together we collect and analyze data on both national and local research questions. As analysis results become available, we present these to local policy-makers, in conjunction with our collaborators. We make recommendations for policy changes that will enhance the program, add new program alternatives, or replace old program alternatives.

### **General Research Questions for All School District Sites in Project 1.1**

The following six research questions are broad questions of interest that we apply to each school district setting. As we conduct the analyses to answer these questions, each school district site serves as the location of an individual study, focused solely on that school district. In the

findings sections of this report, we will present each study separately. Following the five sections discussing the findings and interpretation of each school district's study, we will then present general patterns that have emerged across the five sites, to cross-validate the findings in each individual school district, and compare these findings to our findings in five other school district sites from our research from 1991 to 1996 (Thomas & Collier, 1997). The following are the general research questions addressed in each site. The first three questions describe the data gathered in the initial stages of the research, and the second set of questions pertain to the data analyses conducted in the later stages.

**Initial Stages: Identifying Students, Programs, and Student Outcomes:**

- What are the characteristics of language-minority students upon entry to the school district in terms of primary language, country of origin, first and second language proficiency, amount of previous formal schooling, socioeconomic status as measured by free and reduced lunch, and other student background variables collected by the school district?
- What types of special programs have been provided in each school for English language learners upon entry, and what are the chief distinguishing characteristics of each program, going back in time as many years as the central office staff consider historically meaningful and for which valid data are available?
- What student outcomes are used as measures of academic success for language minority

students, including former English language learners?

### **Later Stages of Data Analyses:**

- After participating in the various special programs, how much time is required for former English language learners to reach educational parity with native-English speakers on the school district's measures of academic success across the curriculum, including nationally normed standardized tests?
- What are the most important student background variables and program implementation variables that affect the long-term school achievement of language-minority students?
- Are there sociocultural/sociolinguistic variables that appear to influence language minority student achievement that vary by school or by geographic region, as identified by school staff?

In addition to these general research questions, the central office bilingual/ESL resource staff and research and evaluation staff of each school district sometimes add specific research questions of local interest that are addressed in the data analyses. Overall, the above research questions focus on the social context of each school system, the characteristics of the language minority students that the school system serves, and the measurement of student outcomes over as many years as can be meaningfully collected, examined by curricular program type that the students are placed in.

### **School District Sites**

An important principle of this research design is that we have examined what exists in current school systems in the U.S., without initially imposing any changes on school practices. After results of the data analyses are presented to the school staff, we do make recommendations for program improvement and we discuss and negotiate these with school district staff. As a result, the policy makers in the school district may choose to implement reforms based on the findings and on our recommendations, but we do not control these matters as in a laboratory experiment.

Each school district participating in this study was promised anonymity, in order to allow them to engage in renewal and reform without undue external interference. Our letter of agreement, signed with each superintendent, states that our participating school districts may identify themselves at any time as well as authorize us to do so, but that, until they do so, we as researchers will report results from our collaborative research only in forms that will preserve their anonymity. In this report, three school districts and one school have decided to self-identify. One school district remains anonymous by their staff's choice.

Also, the participating school systems retain ownership of their data on students, programs, and student outcomes. The researchers have limited rights of access to the data for purposes of collaboratively working with each school district to help them organize, analyze, and interpret existing data collected by the school districts, for the purpose of action-oriented reform from within. However, since the districts own their own data, the researchers may not distribute the data to others. We also provide extensive assurances that we will preserve student anonymity

and will not allow individually identifiable student information to be published.

School districts were chosen through nomination from state education agencies and self-nomination based on the following criteria that we used in our first letter of introduction:

**To be eligible to participate in our research study, a school district should have the following:**

- A district-wide commitment to constructive reform of instruction, backed up by administrative willingness to experiment and to commit resources to evaluation and data collection activities, a willingness to engage in collaborative research to investigate what happens to language minority (LM) students in school in the long term, and active administrative support for the research up to the assistant superintendent level at least;
- A willingness to engage in collaborative research that seeks answers to politically difficult questions, to engage in collaborative development of locally-focused research questions, to collaboratively interpret the research findings with the researchers, and to implement the recommendations that proceed from the collaborative research;
- A willingness to commit to a sustained change process in which the district actively investigates what happens to local LM students in the long term, applies research findings to local decision-making on the most effective program choices for LM students, and actively moves to implement more effective instructional approaches over the next 3-5 years by emphasizing staff development and by providing active support for building administrators' efforts to implement and improve effective programs for LM students;
- Available student-level data stored on magnetic media on recent LM and non-LM student test scores (preferably normal curve equivalents [NCEs] and/or scaled standard scores on norm-referenced tests, but also criterion-referenced tests and performance assessments).

For example, data might be available from years 1997-2001 for high school grades 9-12, from 1994-97 for middle school grades 6-8, and from 1991-94 for elementary grades 3-5;

- Available student-level data on student participation in LM programs in the past (e.g., from 1988 to the present), typically from the central student information system and/or from the Bilingual/ESL office. Data should be either on magnetic media or the school district should be willing to enter it into a computer from paper-based records;
- The district should have local computer capabilities and computer staff sufficient to allow for timely and accurate downloading of existing computerized data from microcomputers or mainframe computers.

**In addition, the following characteristics are desirable in participating school districts:**

- The school district should offer a variety of services to LM students and should be experienced in implementing these services through ongoing staff development;
- The school district should serve a variety of LM populations; districts that serve indigenous ethnolinguistic groups or that provide additional geographic diversity (e.g., rural or under-represented regions) and generalizability are especially desirable;
- In general, mid-to-large size school districts are more desirable than small districts because of larger sample sizes and greater student diversity (but there are exceptions to this);
- The school district should be willing, if needed, to (1) collect additional data (e.g., teacher survey, parent survey, student survey) and (2) convert paper-based student records to

computer-readable form as necessary to address local and national research questions.

**Research sites chosen.** After travel to 26 states to identify school district sites during the year prior to OERI funding and the first year of the grant, 16 sites in 11 states were chosen as best representing the qualifications listed above. Our ultimate goal was to have, by the end of this five-year study, enough longitudinal data from five school districts to report their findings. In order to have extensive well collected data, we knew from previous research experience that it is necessary to collect data from many more sites than required, because many factors influence longitudinal data collection, such as student mobility, change in assessment instruments used by the school district, changes in state policies, new data management systems installed that do not allow retrieval of historical records, and changes in school management that bring about unexpected program changes.

The final five research sites presented in this report were able to make sustained efforts to maintain their programs and data collection systems for the full five years of this study. Their programs were the most consistent and cohesive, and the data management personnel were able to provide the most systematically collected data, and the reform orientation of the school system was maintained throughout the study. Also these five sites represent a purposive sample of some of the major regional contexts of the U.S., demonstrating greatly varied geographical and sociological contexts for schooling language minority students. We are grateful to the four sites (three school districts and one school) that have chosen to self-identify, since that allows for the

richest social description of the context in which the students are schooled. The remaining school district is presented in more general terms, to preserve anonymity.

**Varied locations of research sites.** Regions represented are the northwest, northeast, southeast, and south central U.S. These school sites include two rural school districts in the northeast U.S. on the Canadian border (presented as one study, because of their proximity to each other and their similarity in school population served and programs provided), one inner city school in an urban school district in the northwest, one very large urban school district in the south central U.S., and one middle-sized urban school district in the southeast.

**Linguistic and cultural groups represented.** The primary languages of the students represented in the databases for this study include over 70 languages, but our data analyses in three of the five studies focus on the academic achievement of native Spanish speakers, the largest language minority group in the United States (75 percent of the language minority school-age population). Two of our studies examine the academic achievement of newly arriving immigrants. Two other studies focus on students from ethnolinguistic groups with cultural and linguistic heritages that predate the beginning of the United States—students of French cultural and linguistic roots in the northeast and students of Spanish-speaking heritage in the southwest U.S. The fifth study includes both new immigrants and U.S.-born Hispanic students.

Overall, the data analyses of this research focus on English language learners who begin their schooling with no proficiency in English, but since ELLs do not remain ELLs forever, we refer to them as language minority students (or former ELLs or ESL/bilingual graduates), because

as we follow them across the grades K-12, they make progress in acquiring the English language and they are eventually reclassified as English-proficient. Since all our analyses are long-term, our findings represent former ELLs who are at various stages of proficiency development in English and their primary language, and are gradually reaching grade-level achievement in English.

**Program types represented.** These school districts have well collected data on eight major different program types for English language learners. Each school district provides a different combination of programs. Overall, these school districts provide a very rich picture of variations in schooling for English language learners. The analyses include student outcomes from 90-10 two-way bilingual immersion (or dual language), 50-50 two-way bilingual immersion, 90-10 one-way developmental bilingual education, 50-50 one-way developmental bilingual education, 90-10 transitional bilingual education, 50-50 transitional bilingual education, English as a Second Language (ESL) taught through academic content, and the English mainstream. In this report, we present data analyses that cover student achievement on standardized tests in English and Spanish (when available) for Grades K-5 in three districts and grades K-11 in two districts.

**Student records sample.** The total number of student records collected in the five districts featured in this report is 210,054. One student record includes all the school district records for one student collected during one school year, such as that student's background characteristics (which might include socioeconomic status as measured by free and reduced lunch, level of English proficiency and primary language proficiency upon entry to the school district, and amount of prior formal schooling), the grade level and school program(s) that student

attended, and academic achievement measures administered to that student during the school year. Each school district is different in what data they collect and we found it necessary to customize our generic plan to meet the specific needs and characteristics of each school system.

## **Data Collection**

**Collecting qualitative data.** Qualitative data for this study come from many different sources. To describe the social context for each language group being schooled in a given school system, we collected source documents that include reports and studies conducted by the research and evaluation office and the bilingual/ESL office, program manuals, district-wide reports on student and school demographics, newspaper articles, books that describe the region, professional journal articles, and state legislative policy documents that have an impact on language minority education. We kept detailed records of our interviews with central office administrators, school board members, administrators of the bilingual/ESL programs, principals, teachers, and community members. With each visit to the school district, we collected source documents and conducted interviews with central administrative staff and the bilingual/ESL administrators and resource staff, to analyze current policies and practices.

These source documents and interviews provided important information for analyzing the regional context for educating the language minority groups who attend the schools. Each of the studies for which we have been given permission to identify the school district begins with a section that analyzes some of the historical demographic patterns of culturally and linguistically diverse groups that have settled in that region, followed by a specific focus on the state and then the local context for schooling these diverse groups. Included are some analyses presented from political, economic, historical, sociological, anthropological, and linguistic perspectives.

We also visited some schools and individual classrooms on each visit, to clarify issues in

classroom implementation, but our collaborative researchers—the bilingual/ESL resource staff—were our main source for collecting data on and analyzing general patterns in teachers’ practices. For the smaller school districts where a survey was feasible to use, we collected data from each bilingual/ESL teacher, on a survey instrument that we developed for this study that was designed to categorize their general teaching practices, their teacher certification credentials, and general practices within their school building regarding the languages represented among the student population. This data collection instrument is provided in the appendices of this report. The surveys were administered and verified as accurate by the bilingual/ESL resource staff who regularly visit the teachers’ classrooms and provide staff development assistance as needed.

**Collecting quantitative data.** The following overview outlines some of the important sources for data that we collected from each school system that is stored on magnetic media in machine-readable files, and the process that we went through to prepare this data for the analyses. First, we assisted each school district to identify and gather their existing data from the many sources available in the district: e.g. Registration centers, Language minority/Title VII student databases, Student information system databases, Testing databases, and any other databases collected for state and federal reporting. To start this process, we provided a list of potential variables that could be included in the study, and the bilingual/ESL and research staff of the school district then met with us to jointly determine which variables were important to collect and available in machine-readable form. In some cases, existing databases had to be supplemented with new data, in order to answer research questions of local concern. Second, we

assembled all data records from all sources and linked them by student ID to create year-by-year databases. Third, using relational database software, we compiled multi-year databases from the annual databases, creating an internally consistent data structure across the years.

As each data set arrived, we organized and restructured and cleaned the data to identify any problems in the data sets, in preparation for the initial exploratory, descriptive, and cross-sectional analyses. We also converted each data file from its initial format (FileMaker, dBASE4, Microsoft Access, Microsoft Excel, or fixed-length ASCII records) into the .DBF format of Visual FoxPro, the database package and programming language that we use. The data cleaning and data restructuring stages required much time and effort for several reasons. First, historical data were being collected from each school district, for as many years back as each school district had quality data available, and new data was being collected with each school year, resulting in a large number of annual data sets from each district. Second, since we helped school districts to collect and merge all of their data sources, which were often housed in separate offices, this stage represented a lengthy and complex process of reformatting, merging, and restructuring the data files to achieve compatible data structures and data coding protocols among the various data files originally created by different offices to meet a variety of different needs. We arrived at data structures and coding schemes that allowed us to address and answer each different research question, involving different units of analysis and analytical requirements.

## **Data Analyses**

Once the data sets were restructured for compatibility with the requirements of our research questions, our research analyses proceeded through five stages. Initially, we performed descriptive summaries of each variable, including exploratory data plots and measures of central tendency and variability for each variable studied. After we conferred with the school district staff on any missing data and determined that complete data sets were present for each variable needed to answer the research questions, we used relational database computer programs to create cross-sectional databases that allowed examination of student performance and characteristics at one point in time. Then, we used these cross-sectional databases to create longitudinal databases that followed participating English language learners across the years of their school experiences. We began with longitudinal databases that followed students for at least four years, and then supplemented these with databases of students followed for five years, six years, and so on, up to 12 years, when available. Only students who attended at least 100 days of one school year were included in the analyses.

After analyzing these longitudinal databases separately, we then aggregated them so that all students in a given grade were combined across the years of available data in succeeding waves of students. For example, all those who persisted in the school district for five years (K-4) and who arrived at Grade 4 during either 1989, 1990, 1991, 1992, or 1993 were combined to examine fourth grade performance of all of these five-year cohorts over the past five-year period. Thus, the students who were in Grades K-4 during 1984-89, were combined with the K-4 students from 1985-90, with K-4 students from 1986-91, and so on up to the current school year.

Collectively, these K-4 cohorts formed a “super-cohort” of K-4 students, combined from the current school year back in time for as many years as data were available.

The same analyses were then carried out for the six-year aggregate of fifth graders with six years of schooling. Similar analyses were conducted for each of the remaining school grades. This “layered cohort” approach allowed for full examination of the impact of programs for English language learners (ELLs) on student achievement for the past several years, and allowed for much greater sample sizes to be achieved than are possible in normal longitudinal analyses. Only longitudinal cohorts from the same grade range were combined. We made no use of linked or matched groups containing different students across time. Each cohort consisted of one group of students, followed for as long as they attended school in the district and each “super-cohort” group was analyzed separately.

<b>GRADE</b>	<b>K</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>
all 5 year cohorts					T								
all 6 year cohorts					T								
all 7 year cohorts					T		T						
all 8 year cohorts					T		T						
all 9 year cohorts					T		T		T				
all 10 year cohorts					T		T		T				
all 11 year cohorts					T		T		T				

all 12 year cohorts					T		T		T			T	
all 13 year cohorts					T		T		T			T	

### **The Five Stages of Analysis and the Research Questions for Each Stage**

Our collaborative work with our participating school districts proceeds through five major stages over a period of 3-5 years. These stages, and their intents, evaluative questions, and required data are summarized and discussed in the following pages. This five-stage process initially examines the effects of past programs for English language learners (ELLs), and conducts a needs assessment to determine the size of the achievement gap between ELLs and the native-English-speaking students of the school district. During the initial stages, we work with the local school staff to train teachers in improved implementation of the programs, and help the district set up computerized systems to collect program evaluative data, and we allow the programs to mature to the point that they can be feasibly and validly evaluated.

These five stages provide a template for our research in each school district. As such, the stages generally guide but do not determine our work with the participating school districts. As circumstances and preferences differ among districts and among decision makers within districts, we modify and customize our procedures with each district to better address its characteristics and needs. However, we continue to address the overall concerns of each stage to the greatest degree possible. This flexibility avoids the “one-size-fits-all” problem in which a research study may sacrifice ecological validity in the interest of achieving a “standard” research design. On the

other hand, we adhere to the same general evaluation questions, guidelines for program development, and types of measurement for each district in order to achieve an acceptable level of comparability among our participating districts. These five stages also reflect the program evaluation perspective that appropriate educational inquiry should focus initially on program development, on improving program processes, and on identifying and facilitating theoretically-based factors that should enable eventual program success in eliminating the achievement gap between native-English speakers and English language learners.

Stages 1 and 2 serve to describe and document the context, characteristics, and degree of the achievement gap. Specifically, Stage 1 work documents the achievement gap and brings it to the attention of school district decision makers for a decision as to whether the observed gap will be addressed or ignored. Stage 2, in turn, focuses on the district's English language learners and examines the degree to which they have closed the achievement gap while participating in an ELL program and after they have entered the mainstream curriculum, as broken down by years of program exposure and initial age of students when entering the ELL programs. Stage 3 examines how the achievement gap has developed over time and how it differs among the various ELL programs operated by a school district. It also provides decision-makers with trend data on student achievement by program type that guides further decisions affecting continued program development and improvement. Stage 4 provides a comparison and cross-validation of samples and cohorts in order to improve the generalizability of findings by not limiting the research to only one group of students followed across time. Finally, after the programs have been developed

for several years and allowed to “mature” in terms of their ability to provide the most complete services to ELLs that each program can produce in that school district, Stage 5 addresses the summative questions of relative long-term program effectiveness and the factors that influence it.

## Overview of Stage 1 Evaluation Work

Stage	Major Intent(s)	Primary Evaluative Questions	Data Needed
<b>One</b>	<p><b>A needs assessment</b></p> <p>To document the district’s past achievement outcomes for three mutually exclusive groups of students and to compare the five-year progress of the three groups (i.e., to conduct the Thomas-Collier Test of Equal Educational Opportunity):</p> <p>Group 1: former LEPs (English language learners)</p> <p>Group 2: students who are Language Minority (LM) but never classified as LEP (did not participate in a local LEP program)</p> <p>Group 3: native-English speakers who are not part of groups (1) or (2) above</p>	<p>After five years of appropriate instruction in the district, is there an achievement gap between former LEPs (English language learners) and native-English speakers?</p> <p>Has the achievement gap between former LEPs, LM-but-not-LEPs, and native-English speakers widened, narrowed, or remained the same for the past 5 years?</p> <p>Have groups of special interest (e.g., refusers of ESL services, waived students) widened, narrowed, or maintained their achievement gap in the past 5 years?</p>	<p>downloads of test scores and student classification information from prior years</p> <p>Specifically:</p> <ol style="list-style-type: none"> <li>(1) student ID</li> <li>(2) original student classification</li> <li>(3) date entered school and LEP/ELL program</li> <li>(4) test scores from recent years</li> <li>(5) initial proficiency in English</li> </ol>

**Stage 1: A focused needs assessment.** In Stage 1 analyses, we examine the difference in long-term achievement levels between three mutually exclusive groups: former English language learners (ELLs) who have received local ELL program services, language-minority students who were not classified as ELLs and were not in programs specially designed for ELLs, and non-language-minority native-English speakers. Naming this comparison the Thomas-Collier Test of Equal Educational Opportunity, we have required each of our participating school districts to examine this comparison as a condition of working with us. The Thomas-Collier Test establishes

whether a school district's programs for ELLs are allowing ELLs to reach long-term achievement parity with non-ELLs in the district. It also forces districts to disaggregate the test scores for two frequently combined categories of language minority students—those who have been classified as LEP/ELL and are eligible for services, and those who are not. We have noted in the past that many school districts have “hidden” (intentionally or unintentionally) their English language learners' large achievement gap by reporting together the achievement of ELLs and non-ELLs who are members of language-minority groups. Districts have also focused only on the short-term achievement of these groups, ignoring the fact that achievement gaps continue to develop over time. Stage 1 analyses address this issue by comparing the achievement of language-minority LEP/ELLs served by local programs, language-minority non-LEP/ELLs not served by local programs, and non-language minority native-English speakers. In this way, a clearer and more accurate picture of the impact of local programs on English language learners' achievement emerges. Using the results of these analyses, the district can decide not to address these issues and drop out of our collaborative agreement, or decide to address these issues by continuing on to the successive stages of our joint research. Thus far, no school district has chosen to ignore the findings of Stage 1 analyses and drop out of our collaborative evaluation work.

## Overview of Stage 2 Evaluation Work

Stage	Major Intent	Primary Research Questions	Data Needed
<b>Two</b>	<p>A focus on assessing the achievement of LEP students</p> <p>to document the past and present achievement performance of LEP students (current and former English Language Learners who are in Group 1 from Stage One)</p>	<p>Do current LEP students close the achievement gap with each passing year in the LEP/ELL program?</p> <p>Do former LEP students close the achievement gap while in the regular curriculum?</p> <p>Do older LEP students close the achievement gap differently from younger students?</p>	<p>Additional student information needed:</p> <p>(1) date of birth (2) days attended school each year (3) date of exit from LEP/ELL program</p>

**Stage 2: ELLs’ academic achievement gains by length of residence in the U.S. and age on arrival.** In Stage 2 analyses, we focus on ELLs only and examine their achievement gains over the past 3-5 years. We break down their achievement gains by the students’ length of residency in the U.S. (in the case of immigrants) or number of years of exposure to English. In addition, we break down achievement gains by student age upon entry into LEP/ELL programs or, for immigrants, by their age on arrival. We have found in prior research that ELLs’ abilities to close the achievement gap differ greatly depending on whether they are participating in a LEP/ELL program or have left the ELL program and entered the regular instructional program. Since these prior findings imply that length of ELL program, as well as program quality, are both important factors in closing the large achievement gap, we devote Stage 2 to a thorough investigation of these matters. Thus, these analyses serve to confirm the findings of Stage 1 and

to further explore how the observed achievement gap has developed in the school district. Neither Stage 1 nor Stage 2 examines the particular programs that ELLs received, but Stage 3 does.

### Overview of Stage 3 Evaluation Work

Stage	Major Intent	Primary Research Questions	Data Needed
<b>Three</b>	<p><b>A focus on program “productivity”</b></p> <p>to determine the average annual long-term achievement pre-post gains of former LEP students who participated in various types of programs for LEP students</p>	<p>Which programs allow students to close the achievement gap over time and which do not?</p> <p>Do students in some programs close the achievement gap better or faster than in other programs?</p> <p>For each LEP/ELL program, what is the average sustained achievement gain per year for the past five or more years?</p> <p>How do gap closure rates compare for elementary, middle school, and high school years?</p>	<p>Student program participation data</p> <p>Specifically: (1) program type(s) student received each year</p>

**Stage 3: Achievement gap closure by program type.** In Stage 3 analyses, we examine the degree of achievement gap closure that characterizes each program type that has been offered for ELLs by the school district during the past five years or more. Each program is described by its average rate of gap closure or achievement gain (e.g., 3.7 NCEs per year over a 10 year period) but no attempt is made to control for extraneous variables at this point because only average achievement gain per year is being examined. The research question of interest here is “looking at trend data in a time-series fashion, what has been the average progress of students in each program type, measured as average gain (and degree of achievement gap closure) over the past 3-5

years? Programs in which ELLs have closed the achievement gap are deemed more effective than programs with little or no demonstrated gap closure, independent of the characteristics of participating students or their initial scores at the beginning of the LEP program.

Stage 3 analyses serve several very important functions. First, they provide school district decision-makers with interim, formative information on student achievement that allows a “time-series” comparison of the effectiveness of their various program offerings for ELLs over the past several years. This is a pragmatic response to the political needs of school boards, superintendents, and program administrators to have in-progress interim results from their efforts to design better programs for English language learners. These groups are simply unable and unwilling to wait for years to know whether their efforts to improve ELL education are productive or not.

Second, stage 3 analyses provide useful information to the districts as to which of their past ELL programs have demonstrably closed the achievement gap and which have not. This information can be very enlightening to both administrators and teachers who may be personally convinced of the efficacy of one program type or another, but have never actually examined how student ‘graduates’ of their preferred program really perform in long-term school achievement, as measured by the same tests given to native-English speakers, on-grade-level and in English. The realization that their ‘favorite’ program (whether a type of English-only or English-plus instruction) is not really meeting the needs of their English language learners can serve as a refreshing “reality-check” and as a professional impetus to examine their professional assumptions and change their practices to reflect the characteristics of more demonstrably effective programs. On the other hand, if staff find that their ‘favorite’ program is somewhat effective for ELLs, but can be improved, this serves as an impetus for them to examine their

practices as well, looking for new program strategies and processes that will allow them to improve an already-good program.

Such information is made more useful when conclusions and findings can be confirmed across multiple groups and contexts. Stage 4 addresses these issues of generalizability.

### Overview of Stage 4 Evaluation Work

Stage	Major Intent	Primary Research Questions	Data Needed
<b>Four</b>	<p><b>Enhancing external validity (generalizability) and robustness of findings and conclusions</b></p> <p>Revisit Stages One through Three by:</p> <ul style="list-style-type: none"> <li>(1) adding successive waves of longitudinal cohorts;</li> <li>(2) using cross-validation strategies to compare findings across groups;</li> <li>(3) employing resampling strategies.</li> </ul>	<p>Are the observed between-group and between-program differences in student achievement trends stable and consistent across comparable but different longitudinal cohorts of students during the past 5-10 years?</p> <p>For each program, what are the estimated means and standard deviations of the sampling distribution of findings across comparable grade-groups and cohorts for each program?</p>	<p>Stage 1-3 data for additional student cohorts and additional cross-validation groups</p>

### Stage 4: Increasing sample size by adding more cohorts and re-sampling

**techniques.** In Stage 4, we add as many years of student data and as many longitudinal cohorts of the same students followed over time as are available and reasonable to add, to further increase sample sizes. This addresses the problem of student attrition caused by students leaving the school district, and thus the school districts’ programs for English language learners. In addition, adding more student cohorts and groups provides opportunities for “mini-replication” of findings

from initially-investigated student groups. In principle, this is similar to replicating an initial study, in that a separate but comparable student group is investigated, and findings are compared to those from the initial study. This form of ‘robust’ analysis can add much generalizability to the findings and conclusions of the initial study. In addition, this offers the opportunity to investigate separately any groups whose findings differ significantly from those of similar groups, looking for possible moderator variables or ‘hidden’ variables whose effects on local student achievement had not been previously recognized.

Also, in some instances of Stage 4 work , we use re-sampling techniques (e.g., the bootstrap), a set of statistical methods that yield valid population parameter estimates from local sample statistics to achieve more generalizable estimates of the long-term impact of special programs for ELLs on the English language learners in the school district. Since one of the ultimate objectives of our research and program evaluation efforts is to arrive at useful and valid estimates of the long-term achievement effects of various programs and program strategies for ELLs, re-sampling techniques provide additional insight into what the theoretical “national distribution” of long-term achievement scores would look like for students who had experienced each type of ELL program.

Only after the work of Stages 1-4 has been completed is it appropriate to take up questions of summative long-term program effectiveness in Stage 5. This is the case because it typically takes years to achieve a condition of (1) full development of the ‘school district version’ of each program to its design specifications; (2) full training of the professional staff to understand each program’s instructional features and to deliver these features, and the program, as designed; (3) development of an adequate data-collection system in the school district that will allow on-going analyses of instructionally important variables and student characteristics over time, and not be limited to the typical 1-2 year data collection time frames in which most school districts operate.

## Overview of Stage 5 Evaluation Work

Stage	Major Intent	Primary Research Questions	Data Needed
<b>Five</b>	<p><b>A quasi-experimental focus on LEP achievement by programs with <u>appropriate</u>, best-available control of extraneous variables</b></p> <p>to determine the long-term achievement of LEP students who received selected LEP programs in the past with control of pertinent extraneous variables on the enhanced data sets from Stage Four</p>	<p>With selected extraneous variables controlled using sample selection, blocking, or ANCOVA (if appropriate), are there long-term differences in student achievement among programs?</p>	<p>Student characteristics and other variables to be controlled</p> <p>Specifically:</p> <p>(1) initial grade placement in school</p> <p>(2) free-reduced lunch for each year</p> <p>(3) initial achievement test scores at beginning of schooling in primary language and in English</p> <p>(4) initial proficiency in first language</p> <p>(5) other available student variables from surveys or from district’s student information system</p>

**Stage 5: Repeated-measures ANOVA, Multiple regression analyses and controlling for extraneous variables.** Finally, in Stage 5 of our analyses, we turn to the research question, “Which program is better, when extraneous variables (e.g., initial differences between groups) are controlled?” These analyses are appropriate only after two conditions have been met. First, the programs for English language learners must have “matured” past the point of initial program installation and past the point of resolving “startup bugs.” Second, the programs must have reached a point of full implementation by the school district that is faithful to the specifications and theoretical design features of each of the programs. Otherwise, level and quality of implementation is confounded with program type, resulting in the comparison of

poorly implemented programs of one type with well implemented programs of another type. In order to arrive at valid between-program comparisons, all programs must be meeting their full theoretical potential in terms of implementation, at least to the point that is pragmatically possible within the context of good administrative support and well-trained teachers.

In stage 3, we collect information on program processes as well as on degree and quality of program implementation in each school. We accomplish this by means of surveys directed to each classroom teacher, by interviews with instructional coordinators who observe instruction in the schools for each program, and analyzing any data collected by the school district on how instruction is carried out in each school. These data are added to the data collection system and provide possible variables for use in Stage 5.

### **Quasi-experimental pitfalls**

There are many problems with analyses in Stage 5 when attempting to control for extraneous variables. First, random assignment is almost always not available as a strategy for addressing potential differential selection problems. True random assignment, rather than systematic assignment of students from class lists to programs under the label of ‘random assignment’ is very rarely encountered for very good pragmatic and political reasons. Although some apparently naive researchers have called for randomized studies of ELL program alternatives, school administrators understand the large political difference between randomly assigning students to controversial, politically-sensitive treatment alternatives (e.g., English-only vs. bilingual programs) and assigning them to not-so-controversial alternatives such as slightly

smaller vs. slightly larger classes that were studied in the recent Tennessee STAR evaluation of class size. In the former case, randomly assigning large numbers of students in a school district to program types strongly opposed by the students' parents, a necessary outcome of wide-scale use of random assignment, would amount to political suicide for the responsible school administrators. In the latter case, it was possible to conduct a randomized study in Tennessee because the treatment alternatives were not controversial and because the study was mandated by the state legislature. Thus, those who advocate such large-scale use of random assignment to study ELL programs are, in effect, announcing that they don't really understand the political difference between controversial and not-so-controversial program treatments, and also that they have no actual experience in conducting large-scale data collection and analyses in school districts. It is also worth noting that the most strident advocates of random assignment as a form of "scientific" research on ELL programs may also be those who are interested in reducing funding for such research by imposing funding conditions that are virtually impossible to meet in the typical school district.

Second, even in the rare cases when random assignment of students to different program alternatives is possible (e.g., it is illegal in the U.S. to randomly assign limited-English-proficient [LEP] students to no program treatment, so true "no-treatment" control groups are very difficult to arrange), we have observed that its effects in initially equated groups begin to deteriorate rapidly in a program that lasts more than about 2-3 years. This increasing group inequality over extended time periods is caused by the fact that students don't leave school for random reasons,

either between programs or within programs, even when they have been randomly assigned to groups initially. This is especially true if the groups are of typical classroom size (15-30 students per group) because random assignment is a large group strategy and can often yield quite unequal groups when employed with small samples.

The interested researcher may verify this by taking a large sample of student records, randomly assigning the students to two arbitrary groups, and then comparing the groups on a fixed variable both initially and then again 4-5 years later, after substantial attrition has taken place in both groups. In many cases, the initially equated groups (e.g., average ages are the same in each group) are no longer equated after several years (i.e., average ages are significantly different in the two groups), because of differential student attrition in the two groups from non-random causes. Thus, we have found that random assignment works consistently only in short-term studies. However, in the short term of 1-2 years, small annual and cumulative effect sizes may not be detectable by statistical significance tests of appropriate power, until they reach values equivalent to .20-.30 standard deviations. Since most programs for ELLs have small annual effect sizes, this requires at least five years, thus making long-term studies mandatory.

A third problem is that the “scientific” use of analysis of covariance (ANCOVA) to ‘equate’ unequal groups after the fact is fraught with problems associated with violation of its necessary assumptions of linear relationship between covariate and dependent variable and between covariates, reliability of covariates, and homogeneity of regression, in addition to the usual ANOVA assumptions of normality and homogeneity of variance. The homogeneity of

regression assumption must be tested explicitly for each ANCOVA or one runs the grave risk of either over-adjusting or under-adjusting the group means. If either of these happens, one has essentially removed real differences between groups or created artificial differences between groups. Either way, the legitimate comparison of ‘comparable’ students in different programs is quite invalidated from that point on. For all of these reasons, the use of true random assignment in evaluation of programs for English language learners is virtually impossible, despite naive calls for this by some researchers and politicians.

Therefore, in our stage 5 work, instead of random assignment, we use ANCOVA when its assumptions are met, and blocking in other circumstances. One can use blocking to create new independent variables (that might have been used as covariates) that are crossed with the independent variable of interest, Program Type. In this way, variation due to the potential covariate is removed and assessed separately as another independent variable and the effect of Program Type is analyzed as in a typical ANOVA. A significant interaction between Program Type and blocked independent variable indicates that the homogeneity of regression assumption would have been violated in an analysis of covariance, thus invalidating it. In addition, blocking is advantageous because it does not require the satisfaction of ANCOVA-type assumptions, and its power approaches that of ANCOVA when there are three or more groups defined in the blocked variable. In many cases, simply analyzing separately the groups defined by a blocked variable (e.g., separate longitudinal analysis of student achievement gains by program type for students of low, mid, and high socioeconomic status [SES]) achieves results that are quite useful for decision-making, without directly adjusting, often

inappropriately, the dependent variable for the covariate SES, as in ANCOVA. If a consistent pattern of findings emerges (e.g., low SES students always score higher when in a two-way developmental bilingual program than do comparably low SES students in ESL Pullout programs), the researcher's confidence in the validity of the findings is bolstered to the point of utility in decision-making, without the use of random assignment, ANCOVA, or other pragmatically non-useful strategies.

### **Collaborative Interpretation of Data Analyses**

When the data analyses from Stages 1-5 are completed, we return to the school districts for collaborative interpretation of the results with the bilingual/ESL central office staff and research and evaluation staff. Sometimes this leads to the decision to collect additional data, or to reanalyze the data, focusing on new or revised research questions of local interest. The process is cyclical and ongoing, and leads to changes in school policies and programs, collaboratively agreed upon by all decision makers in the school district. If the school districts wish to continue in this cyclical reform process by continuing to grant us access to their student data and test scores, we are presented with the opportunity to continue to engage with them in a “recycling” to earlier stages of our five-stage research process, and continued collaboration in their ELL program renewal efforts.