5-2004

An Analysis of Bilingual Programs in the Context of a Schoolwide Reading Program

Jonathan A. Stewart
Utah State University

AN ANALYSIS OF BILINGUAL PROGRAMS IN THE CONTEXT OF A SCHOOLWIDE READING PROGRAM

by

Jonathan A. Stewart

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology
(School Psychology)

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

2004

ABSTRACT

An Analysis of Bilingual Programs in the Context

of a Schoolwide Reading Program

by

Jonathan A. Stewart, Master of Science

Utah State University, 2004

Major Professor: Donna Gilbertson, PhD
Department: Psychology

There has been much controversy over the effectiveness of bilingual education in helping English language learning (ELL) students to become successful students. One variable overlooked in this literature has been the use of effective instruction in these programs. This investigation compared students in a schoolwide reading program that utilizes research-based practices, Success for All (SFA) and its Spanish counterpart Éxito Para Todos (EPT). Three groups of third-grade students were compared at 8-week intervals throughout the school year: English-speaking students in SFA, ELL (English language learning) students in SFA with ESL (English as a Second Language), and ELL students in EPT. All three groups experienced gains over the school year, with the gap between the EPT and SFA only groups narrowed and no statistically significant differences discovered between the EPT and SFA + ESL groups.

(81pages)

CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Since the passage of the Bilingual Education Act in 1968, effective practices in bilingual education have been a constant concern in education and political spheres. Court decisions in several court cases (*Castañeda vs. Pickard,* 1981; *Lau vs. Nichols,* 1974*)* have mandated that schools should not deny English language learning (ELL) students educational experiences based on the student's language of origin. Federal regulations require school personnel to take "appropriate action" to provide special assistance to facilitate English competence in a manner that will enhance academic performance throughout a student's school career. Such special assistance should be based on sound educational theory, be adequately implemented, and be periodically evaluated. Information obtained from the periodic evaluations must show that the program is producing desirable outcomes or, if not, evaluation information must be used to make program modifications (*Castañeda v. Pickard,* 1981).

By legislative and judicial mandate, school districts must provide special support for ELL students, but are given no direction or guidance as to how this support must be implemented. As a result, a number of different types of bilingual education programs have been added to school services to replace the historical "sink or swim" approach initially implemented, such as English as a Second Language (ESL) programs, English immersion programs, and transitional and maintenance bilingual education (Gersten & Woodward, 1994; Ochoa, Rivera, & Powell, 1997). Due to serious methodological flaws in evaluation studies and inconsistent implementation of programs' components, research

in this area has not clearly indicated the approach to bilingual education that best fosters academic progress for ELL students (Greene, 1998; Lam, 1992; Rossell, 1998; Rossell & Baker, 1996; Thomas & Collier, 1997; Willig, 1985).

The search for an effective bilingual program for bilingual students is likely to escalate with the addition of up to one million ESL students expected in schools over the next decade (Council of Chief State School Officers, 1998). The term ESL is an older term for ELL in this literature and is used interchangeably when referring to students. Spanish-speaking students are a particular focus of educators because they currently comprise approximately 75% of the ELL population. Moreover, projections by the U.S. Census Bureau estimate the Hispanic population in the United States will be 55 million, or 17% of the population in 2020 and 190 million, or 33% of the population by the year 2100 (U.S. Census, 2000).

With such dramatic increases in the Hispanic population, effective and efficient practices are needed to help eliminate language barriers and promote academic success. Currently, academic achievement scores for Hispanic students are significantly below the national average with a substantial gap between achievement outcomes for ELL students and native English-speaking students that increases rather than decreases as ELL students progress through school (Goldenberg & Gallimore, 1991). Hispanic students are two and a half times less likely to be at or above grade level for reading and math in fourth and eighth grade on national standardized testing compared to the national average (National Center for Educational Statistics, 1999, 2000a). Thus, ELL students are required to learn more material at a higher rate than classmates who are already performing at a higher level, making it hard to "catch up."

A key predictor of academic success and school adjustment is the ability to read in English at an early age, thus, it is just as critical for Hispanic students to learn to read in their early years as it is for English-speaking students. During this critical period, educators vacillate between delegating academic time and resources to teach students reading in English or to enhance reading in the student's primary language. Research studies have focused on comparing these two educational approaches in an attempt to determine which strategy will lead to the best long-term outcome. Findings have been mixed, mostly due to variability in methodological rigor employed by researchers (Greene, 1998; Lam, 1992; Rossell & Baker, 1996; Willig, 1985). The few researchers using sound methods generally suggest that students who can proficiently read in their primary language tend to learn to read in English at a faster rate than students who are not proficient in their native language (e.g., Ramirez, 1992).

To date, few researchers have examined the effectiveness of additional quality instructional practices in bilingual programs. Several investigators have identified critical instructional variables (academic engagement, immediate feedback, and progress monitoring) that facilitate the development of classroom reading skills for native English-speaking students (Greenwood, 1996). Despite these findings, teachers provide fewer effective instructional variables to low socioeconomic status (SES) students (Greenwood). For example, Greenwood (1991) recorded low SES students in second to fourth grades receiving substantially fewer response opportunities and obtaining lower standardized academic scores than high SES students in classroom observations. Because 80% of the ELL student population is below the poverty level (Baker & Hakuta, 1997),

one could speculate that this phenomenon is occurring with a majority of ELL students as well.

Several schoolwide reform programs, such as the reading program Success for All (SFA), have emerged to systematically increase the implementation of effective instructional variables (Slavin, 1995). Slavin and Madden (1999a), investigated the effects of SFA along with bilingual programs and reported effect sizes ranging from 0.2 - 1.0 from various schools and districts on reading outcome measures. Variables that affect the wide range in effective sizes across studies, however, have yet to be determined.

A key program issue is whether effective teaching practices used with ELL students narrow achievement gaps between ELL and native English-speaking students and if this progress can be made in an English-only program with ESL support or whether a more intensive bilingual program is needed for student success. Further, few researchers have empirically evaluated reading growth trajectories for bilingual students using frequent data points throughout a school year. An evaluation of growth rates with an intensive instructional program will help determine what type of bilingual program would facilitate learning at similar or greater learning rates than English-speaking students (Fuchs & Fuchs, 1998).

The purpose of this study is to evaluate and compare the effect of a systematic instructional program, SFA, on reading level and rate performance for language minority students and native English-speaking third grade students. All English-speaking students and one group of ELL students will be provided with the SFA program in first through third grade, with this ELL group being provided with ESL support. A second group of ELL students will be provided with Éxito Para Todos (EPT), a reading program in their

native language of Spanish. ELL students in the EPT and ESL groups are then compared in third grade with the native English-speaking students to determine which program has a greater positive effect on English reading ability over time.

CHAPTER II

REVIEW OF LITERATURE

In response to the increasing number of bilingual students and their lack of progress, the effectiveness of bilingual education has been the subject of numerous studies (Rossell & Baker, 1996; Willig 1985). Based on different theories of language acquisition, various types of bilingual programs have been developed and investigated. However, due to conflicting findings in the literature regarding the effectiveness of bilingual education on academic outcomes, empirical research fails to indicate one broad approach that would best promote learning (Gersten & Woodward, 1994).

Researchers who have analyzed the body of bilingual education research suggest that there are several reasons for the mixed findings in this area (Greene, 1998; Lam, 1992; Rossell, 1998; Rossell & Baker, 1996; Thomas & Collier, 1997; Willig, 1985). First, investigators often do not clearly describe what type of program was used or how proficiently or consistently teachers in the classroom implemented the program. Hence, variance in research outcomes often reflects differences in program components not directly analyzed, or differences are due to differences in the degree to which all components are implemented. Second, inconsistent findings may be due to methodological flaws that obscure the actual effectiveness of bilingual education. Specifically, small sample size, attrition, nonrandomized samples, inappropriate statistics, lack of equivalent group comparisons, weak experimental control, poorly defined program protocols, and use of measurements with poor psychometric properties are common methodological limitations THAT hinder decisive conclusions regarding the

impact of bilingual education on academic progress. In fact, when various reviewers attempted to summarize the research evidence on bilingual programs by excluding studies with weak research methods, an average of only 10% of studies reviewed were accepted for further analysis of program effectiveness (Lam). Third, it is extremely difficult for researchers to experimentally measure and control the vast number of background characteristics that vary among individual ELL students. Despite these limitations, it appears that students who receive some level of bilingual education generally obtain greater achievement levels than students who receive English-only programs (Greene; Ramirez, 1992; Thomas & Collier, 1997, 2002; Willig).

This section will summarize the empirically based research on bilingual education.

Types of Bilingual Programs

Through efforts to improve educational outcomes for ELL students, different types of bilingual programs have been established to promote English-language skills. One major difference among programs is the wide variability in the number of components added to strengthen primary language skills, increase English proficiency, and remediate overall language skills. In order to better understand the full range of established bilingual education programs, the purpose of individual programs commonly employed in school will be summarized in this section. There are three main models of bilingual education that have been empirically evaluated: immersion, transitional, and maintenance (Gersten & Woodward, 1994; Ochoa et al., 1997). Other strategies, such as

ESL pullout services and "submersion" are not well researched but are approaches that are used or have been used in schools to promote English language skills.

No Intervention or Submersion

The term submersion describes the absence of any native language instruction or other kind of intervention for students who have limited English proficiency as the result of having another primary language. Historically, schools commonly employed this approach until legislative actions first addressed the educational needs of students with limited English proficiency. Currently, this is a strategy that is not legally used, as federal regulations and court cases (*Castañeda vs. Pickard, 1981; Lau vs. Nichols, 1974)* mandate appropriate testing and programming for children with limited English proficiency.

Immersion

Immersion, also referred to as structured immersion, is an approach designed to teach a child in an English setting with very limited native language instruction. The instruction is geared to increase the child's language proficiency in English and students are kept with their classmates. The teacher is fluent in English and has some skill in the child's primary language. Instruction is given in English only, teachers make special modifications to simplify their English, and the native language is used on rare occasions when necessary to complete a task (Gersten & Woodward, 1994; Rossell & Baker, 1996). Canadian educators have used the immersion model for many years and investigators have demonstrated increases in IQ, divergent thinking skills, and increased competence in the second language when compared to matched control groups (Lambert, 1992).

ESL Pull-Out

English as a Second Language (ESL) pull-out services are a form of immersion and are commonly used in schools today. Students are in regular classes with English instruction for academics and pulled out during class for a specialized English curriculum structured to facilitate the acquisition of English. Generally, students are instructed for a half hour to an hour each day by a teacher certified to teach ESL. Teachers in this type of program are not necessarily fluent in the primary language of the student (Rossell & Baker, 1996). Native language instruction is minimal, if used at all. This type of program has been primarily included in studies in which investigators compared the effects of different types of bilingual programs on student academic achievement. The results of these comparison studies have suggested that, while ESL pull-out is more effective than no bilingual services, it typically is less effective than other bilingual programs (Willig, 1985).

Transitional

Transitional or "early-exit" strategies focus on teaching children in their native language in early grades after which students gradually receive increased curriculum in English until the student has "transitioned" into complete English instruction. The time frame for this varies, but generally three years is the target to fully transition a child (Gersten & Woodward, 1995; Rossell & Baker, 1996; Thomas & Collier, 1997). Students receive some degree of instruction, mostly in academic content areas, in their native language and progressively more English introduced into instruction. The goal of this program is to transition to the English language instruction of curriculum as rapidly as

possible. Goldenberg and Gallimore (1991) demonstrated an increase in reading achievement for second and third grade students at the 25th - 35th percentiles, respectively, to about the 60th percentile for both grades at each subsequent grade over a 5-year period after implementation of a transitional program.

Maintenance

Maintenance programs, also known as developmental bilingual education, enrichment bilingual education, or "late-exit" strategies, focus on teaching curriculum material to students in their native language. Students receive instruction geared toward English acquisition, however the program continues to teach a student in the native tongue in academic areas until the student has demonstrated an adequate grasp of English in conjunction with proficiency in academic subjects in his/her native language (Gersten & Woodward, 1994; Ochoa et al., 1997; Rossell & Baker, 1996; Thomas & Collier, 1997). The long-term goal is to develop and maintain cognitive academic language proficiency (CALP) in both languages. Ramirez (1992) found maintenance bilingual education to be the most effective strategy in their large-scale study comparing the effectiveness of different approaches to bilingual education.

Comparison Analysis of Current Bilingual Programs

There have been many attempts by researchers to determine both the effectiveness of bilingual education and which type of bilingual program better impacts long-term academic achievement. The purpose of this section is to summarize the recent major reviews and studies that compare the effectiveness of different bilingual education

programs. This is divided into two sections, longitudinal studies and large-scale reviews/meta-analyses.

Meta-Analyses on the Effectiveness of Bilingual Education

One of the first attempts to get a broad understanding of effective bilingual education was a meta-analysis conducted by Willig (1985). Meta-analytic techniques were applied to studies conducted before 1981. Modest positive effects (mean effect size = .63) were shown for bilingual education groups versus a control group. The author also quantitatively noted that results on bilingual education were generally obtained using poor methodology and positive effects for bilingual education became evident only after using statistical controls for methodological flaws.

A review was conducted by Rossell and Baker (1996) to examine differences between different types of bilingual program. Rossell and Baker examined over 500 program evaluations and journal articles on bilingual education between 1900 and 1995 to judge whether transitional bilingual education is superior to other forms of bilingual education. Studies were then included as methodologically acceptable based on four methodological characteristics: random assignment to treatment and control, nonrandom assignment that matched students in treatment and control groups on relevant variables that influence academic performance, comparison group of limited English proficient students with same ethnicity and language background, and outcome measures using normal curve equivalents (NCE's), raw scores, scaled scores, or percentiles. The application of these criteria left 72 methodologically acceptable studies for consideration.

Given that transitional bilingual education programs was the most common model used, Rossell and Baker (1996) compared transitional programs to submersion, ESL pull-out, structured immersion, and maintenance bilingual education. For the purposes of this comparison, a bilingual program was considered to be effective if students obtained higher reading performance scores that were statistically significant in a bilingual program than students in a comparison bilingual program or in nonbilingual classes. The authors then calculated the percentage of studies that reported greater academic gains in transitional education when compared to other types of programs.

Their results indicated that a transitional bilingual education program was more effective than a maintenance bilingual program but only one study compared these types of programs. In contrast, transitional bilingual education was found to more effectively increase reading performance between the range of 0 - 22% when compared to either submersion, ESL, or structured immersion programs. These results suggested that educating a child in transitional bilingual education may be less effective than other programming options but more effective than a maintenance program.

Using the studies in the Rossell and Baker (1996) review, Greene (1998) further expanded on these results by reexamining the group of 75 "methodologically acceptable" studies and applying meta-analytic techniques to calculate effect sizes for treatment effects. To elucidate differences between types of programs, Greene simply examined the outcome differences between programs that incorporated some form of native language instruction in the teaching process with programs where instructors only taught in English. Of the 75 studies initially included, only 11 were used in the final analysis due to additional methodological concerns in the studies. Reasons for exclusion of studies

included redundancy (two reports of the same study), failure of the authors to report statistics for a meta-analysis, investigators did not directly evaluate bilingual education, lack of an appropriate control group revealed upon further analysis, and investigators evaluated a program for less than a year. In a secondary analysis, Greene analyzed five studies that utilized random assignment in their research design.

In this investigation, Greene (1998) reported the average mean effect sizes (Hedge's *g*) between native language instruction and English-only instruction to be .21 for reading in English, .12 for math (in English), and .74 for Spanish reading. That is, on the average, bilingual education had a small, but positive effect on academic progress for ELL subjects when compared to ELL students in the nonbilingual education groups on English language tests. When looking at studies that incorporated random assignment to bilingual treatment and English-only control, effect sizes were even higher. The average effect size was .24 in overall English content, .41 reading (in English), .15 math (in English), and .92 in Spanish reading. The effects were not very meaningful in math, small in English reading studies with a control group, medium in English reading for studies with a control group, and a large effect size for Spanish reading.

These studies showed mixed results. It is interesting to note that the results varied by the type of method used in the study. In other words, the two studies that utilized meta-analytic techniques showed positive effects for bilingual programs and the study that used a vote counting method showed little evidence of effectiveness for bilingual programs. Another interesting aspect is that 64 of the 75 studies used by Rossell and Baker (1996) for their review did not meet their own standards for methodological rigor when Greene (1998) then tried to take those same studies and apply meta-analytic

techniques. Thus it would appear that, of the reviews of the pool of research studies used by both Rossell and Baker and Greene, Greene's appeared to have done a better job of analysis of these studies. With this factored in, it would appear that the reviews and meta-analyses also showed positive effects for bilingual programs that included native language instruction.

In summary of both types of research, findings indicated that some level of bilingual services had been found to be more effective than English-only services and many studies suggested that ELL students further benefitted from some degree of instruction given in their primary language (August & Hakuta, 1997). Although no ideal bilingual education program was found, several general findings suggested major challenges to the effectiveness in all programs when attempting to promote school success. For example, program implementation was not addressed in these reviews and may hinder potential beneficial outcomes. Yet even with adequate implementation, findings from several studies suggested that oral English proficiency (i.e., conversational language) takes approximately two years, while academic English proficiency (i.e., language use in academic context) can take 4 - 7 years (Cummins, 1999; Hakuta, Butler, & Witt, 2000; Thomas & Collier, 1997). Possibly due to this delay in academic English proficiency, an achievement gap between native English-speakers and bilingual students was common; and this gap increased rather than decreased throughout the school years. Thus, a primary challenge to bilingual education programs is the prevention of diverging trajectories in academic growth between English-speaking students and ELL students throughout a student's school experience.

Longitudinal Studies

One of the major longitudinal studies completed in this area was conducted by Ramirez (1992) for the Department of Education. Their purpose was to examine immersion, transitional, and maintenance bilingual education programs to discover the relative effectiveness of each approach on student performance. Students demonstrated progress in all of these programs, with programs incorporating a higher level of primary language instruction having greater academic success. The results of this study showed no significant difference between the immersion and transitional bilingual educational programs, indicating that children's academic progress increased at about the same rates in both programs. However, for the maintenance bilingual programs, there was a significant gain in English reading, language, and math skills. These growth rates were even higher for sites where a sizeable portion (approximately 40% or higher) of the instruction was given in the student's primary native language.

Thomas and Collier (1997) further examined differences in bilingual education over a wider range of grade levels. In this study, the authors conducted a longitudinal descriptive cohort analysis. They evaluated growth trends in academic achievement for students receiving a given type of bilingual service in cohorts of 4 - 8 years, that is, tracking the students in the same grade for 4 - 8 years. Results of long-term cohorts (e.g., 8 years) were weighted and combined with data from shorter cohorts (e.g., 4 - 7 years) and the trend for academic achievement was evaluated. These cohorts covered students from 1st to 12th grade. The trend line for each program started at the same level in first grade, a little over one standard deviation below the mean on standardized testing. All

students were either in a maintenance bilingual education program, a transitional bilingual education program, or a traditional ESL program.

According to their analyses, students in maintenance bilingual education scored at or above the mean of native English-speaking students in high school on standardized tests. Students in transitional bilingual education, however, scored approximately one half standard deviation below the mean at high school, Students receiving ESL services without academic content taught in their native language scored a little more than one standard deviation below the mean. In summary, their report suggested that when a greater amount of instruction was given in an English language learner's primary language, long-term educational outcome was enhanced for the student.

The Thomas and Collier (1997) report has been critiqued as having major methodological flaws primarily due to lack of statistical analysis and a lack of detail about the cohorts used for analysis (Rossell, 1998). In a follow-up to their original study, Thomas and Collier (2002) replicated their original findings with a different data set that included more detailed statistical analyses. However, while the achievement of these students was compared after a certain time period in bilingual programs, any differences between the students initially were not statistically controlled for and the same students were not followed over time. Because students were obtained from five districts, the authors could not control for differences in district instructional practices and testing.

Gersten and Woodward (1994) conducted a longitudinal study that evaluated the effects of transitional and immersion bilingual programs on the Iowa Test of Basic Skills within one district for students in fourth through seventh grade. The data indicated that academic gains were significantly greater in fifth grade in the immersion program when

compared to the transitional program. This difference in academic gains between the two groups decreased over time such that both groups of students were performing at the mean 24th percentile for the immersion group and 21st percentile for the transitional group.

The results of these studies would indicate that students are learning in all bilingual programs; however, achievement gains varied between the different type of bilingual program. Specifically, students were shown to have the highest overall achievement in the long-term when receiving a program that had a Spanish language component for a significant amount of time (maintenance) or a significant amount of native language instruction. Over time, these effects between programs are small.

Effective Instructional Variables

The type of bilingual program used is not the only variable that will impact academic achievement. As recognized by the study conducted by the National Research Council of the National Academy of Sciences (Snow, Burns, & Griffin, 1998), a standardized bilingual education program designed to increase language proficiency is only one part of the solution for remedying deficits in academic performance. A brief overview of instructional variables and practices will be given.

Snow et al. (1998), in their research findings, indicated that many Hispanic children with limited English proficiency who were instructed and tested in Spanish demonstrate reading difficulties in Spanish and early successful reading performance highly predicted successful academic achievement and school adjustment. This finding suggested that linguistic differences are not solely responsible for poor performance.

Other risk factors included poorly educated parents, home literacy background, low family income, poor school quality, and differences in educational values. According to Snow et al., these variables accounted for low levels of academic achievement within English-speaking children as well as Hispanic students.

Fortunately, there are critical instructional factors that influence student success in reading that can be altered by educators to help promote the performance of students who are at-risk (August & Hakuta, 1997). Some of the key variables include: opportunities to respond (Gersten & Baker, 2000; Schmidt, Rozendal, & Greenman, 2002), cooperative learning (Calderón, Hertz-Lazarowitz, & Slavin, 1998), delivering instruction at the child's level (Gersten & Baker), frequent evaluation (Carter & Chatfield, 1986; Gersten & Baker), and feedback (Walberg, 1992). In addition, there are a number of variables that effectively enhance academic progress and can be manipulated by school administrators. For example, low student-teacher ratio, frequent staff training, and teacher support teams all enhance educational achievement (Greenwood, Delquadri, & Bulgren, 1993; Mosteller, Light, & Sachs, 1996; Stone, 1998; Stringfield & Teddlie, 1991).

Investigators using classroom observational studies to determine performance have found a wide difference in the frequency with which these strategies are used in individual classrooms (Fletcher, Bos, & Johnson; 1999; Greenwood, 1991; Turner & Meyer, 2000). Greenwood et al. (1993) examined the potential effect of several factors that may increase teacher's use of empirically supported practices in their classroom instruction. They found that the implementation of an administration-adoption schoolwide support model with effective instruction practices resulted in an increase in

teacher participation in the program from 30% at baseline to 82% in Year 2. Carter and Chatfield (1986) noted in a case study of a school district with a "successful bilingual program" that one component that added to the "positive school social climate" was a well-defined district curriculum with emphasis on proven instructional practices such as cooperative learning. This recent trend of developing schoolwide instructional programs has lead to incorporation of many of these practices within the curriculum. However, limited research has been conducted to investigate the relationship between effective instructional, schoolwide practices, bilingual education programs, and the effect on student outcomes (Ochoa & Perez, 1995).

Success for All

With the recent nationwide concern about poor reading performance, a plethora of programs have appeared with the intention of improving reading, especially among at-risk students. One is SFA, a program developed by Slavin and Madden (1995, 1999a, 1999b), and their colleagues at Johns Hopkins University. Success for All began as a research-based reading program focusing on early intervention and prevention of reading difficulties. The program specifically targets traditionally low-achieving children to help them attain successful reading performance in elementary school.

Effective Instructional Elements Within the Program

The SFA program has several facets incorporated into the program, many of which are based on research on effective instructional strategies. First, the program includes a number of components designed to increase practice opportunities to facilitate

reading fluency and comprehension (August & Hakuta, 1997). These include partner reading, story-related writing, guided group readings, story retelling, and say-spell-say (i.e., say a word, spell it, then repeat the word). The class sizes are also reduced to 18 - 22 students per classroom to increase the potential number of teacher-student interactions. Students who are struggling are additionally assigned to a certified adult tutor with one-on-one time for 20 minutes daily. Second, by regrouping students based on skill level, students learn reading skills at their instructional level (Adams, 1990; Slavin, 1995). For example, a third-grade student who is reading at a second-grade level may be instructed on second-grade materials with other second- or first-grade students reading at the second-grade level.

Third, a major emphasis is on prereading skills including phonological processing, which has been established as the greatest predictor of reading ability (Adams, 1990). Fourth, another important component of this program includes frequent progress monitoring. Reading assessments are conducted every 8 weeks and students advance, maintain, or remediate based on assessment data and teacher evaluation. Fifth, teachers receive intensive training, provided via classroom observations and feedback, while the teacher implements principles and skills in the classroom (Slavin & Madden, 1999b). Finally, faculty support is provided as faculty work together with an on-site SFA coordinator during weekly or monthly meetings to discuss how to best help children who need additional support or modifications. Family support teams are also in place to serve the social needs of the student and their family, and build a bridge between the school and family to increase parental involvement in their child's education.

Program Effectiveness

Overall, research-based evidence suggests that SFA increases reading performance for students in elementary grades. Specifically, SFA has been demonstrated to be effective in raising reading scores relative to "control" programs in 11 school districts, with the largest gains reported for the lowest 25% of students in their grade (Slavin & Madden, 1999b). A recent review of schoolwide reform programs by the American Institutes for Research indicated SFA to be one of three programs showing "strong" evidence for positive academic gains by students (Herman et al., 1999).

To accommodate programs that include bilingual students, SFA authors have also developed a version of their program, called EPT, using the same instructional methods to teach students to read Spanish (Slavin & Madden, 1995; Slavin & Madden, 1999a). Recently, EPT has been systematically studied in Philadelphia, California, and Houston school districts, as reported in Slavin and Madden (1999a). Student progress was evaluated by gains made on the Spanish and English versions of the Woodcock Language Proficiency Battery (WLPB; Woodcock, 1984). Three subtests were the basis for comparison: word identification, word attack, and passage comprehension.

Philadelphia. The effect of an EPT program was studied in two schools; one in which EPT was implemented and another which used an immersion approach to teaching reading to bilingual students. Participants were initially matched on criteria such as schoolwide Hispanic enrollment, percent of students receiving free lunch, and schoolwide mean percentile in reading. The two schools implemented EPT programs for participating students from first grade to third grade. As predicted, the effect sizes were large in favor of the EPT group for all three WLPB subtests given in the Spanish

language (median effect size = +2.62). English language scores on the WLPB subtests were also positively affected (median effect size = +0.21) to a much smaller degree. However, these results were limited due to small sample size (total $N = 40$).

California. For a 3-year period, reading performance data from three schools in California were compared among three programs for Spanish-dominant students: Spanish bilingual instruction, English instruction, and EPT. All schools were incorporating SFA procedures in the English-language reading classes. As part of the EPT and the Spanish-speaking instructional program, children were transitioned out of the EPT or Spanish-speaking instruction and into the English-language SFA program when they demonstrated readiness for English.

When the students' reading performance in English was assessed at each grade level, the EPT students' mean score was higher than the mean of students in the other programs, although a statistical analysis of significant differences between groups was not reported. However, effect sizes showed a diminishing trend in gains for the EPT students from an effect size of +1.03 in first grade to +0.44 in second grade to +0.23 in third grade. The diminished effect sizes in the higher grades were most likely influenced by the difference in the number and type of students who still remained in the EPT program. That is, students were transitioned out of the EPT program at twice the rate of students in the Spanish-speaking program. Therefore, many students initially participating in the study were no longer included as part of the EPT cohort when tested in the higher grades. Student academic progress in subsequent school years after termination of the EPT program was not investigated in this study.

Houston. Program effectiveness for academic achievement was evaluated for first graders from 20 schools using EPT compared with 10 matched schools in the district that used another form of Spanish bilingual instruction. EPT schools were rated as either high-, medium-, or low-implementation of EPT. None of the schools were rated as having high-implementation and the total number of subjects was roughly evenly divided between medium- and low-implementation. Results indicated the students in the EPT program obtained higher scores on the Spanish version of the WLPB than schools with the other Spanish bilingual instruction. The mean effect size for the medium-implementation schools was 0.24 and for the low-implementation schools 0.17 at posttest.

Summary

Since the passage of the Bilingual Act in 1968, schools and researchers have continually struggled with the establishment of an effective program that would best meet the needs of ELL students. Bilingual education has many different forms, including immersion, transitional, maintenance, and ESL pull-out. Disagreement exists among researchers regarding the effectiveness of bilingual education for ELL students as well as which model is most effective. Even researchers evaluating identical groups of empirical studies come to different conclusions about whether studies are "methodologically acceptable" or show positive effects for bilingual education (Greene, 1998; Rossell & Baker, 1996).

Quality of instruction is frequently mentioned as an important factor in current studies evaluating bilingual programs, but few investigators directly examined its effects

(Gersten & Woodward, 1994; Greene, 1998; Krashen, 1998; Rossell & Baker, 1996; Thomas & Collier, 1997). The SFA program is one research-based reading program with evidence supporting its efficacy (Slavin & Madden, 1999b) for promoting English-reading performance with the largest gains for at-risk students. The primary focus of SFA is to incorporate effective principles of instruction including class-size reduction, a structured approach to reading, increasing reading time, individualized tutoring, and use of homogeneous grouping.

Research has also shown modest evidence that the SFA adaptation to teach the Spanish language, EPT, is effective in teaching Spanish literacy skills that may generalize to the acquisition of English literacy skills (Slavin & Madden, 1995, 1999a). However, limited support is provided for the effect of EPT on English reading and how growth progressed over time as compared to native English-speaking students. Due to methodological limitations of previous research, a valid estimate of program effectiveness requires further study, replication, and verification. Moreover, investigators of studies failed to evaluate performance of students after the bilingual program was terminated for students who met an exit criterion based on the attainment of a specific level of English proficiency within the same district receiving controlling for instructional programs. Success based on English proficiency may not indicate that these students would continue to make academic gains when receiving instruction in English.

Due to highly variable ELL student characteristics, it is likely that continued success would vary substantially among students when special assistance is reduced or terminated. Systematic progress monitoring may indicate how well a student progresses after termination of special services and which students may need additional instructional

services. A measure taken at a fixed point in time would serve to compare an ELL student's level of performance with native English-speaking peers. However, an analysis of rate of growth would be necessary to understand whether or not ELL students are demonstrating a similar capacity to benefit from English language instruction as native English-speaking students. If a child is performing at a lower level than peers but maintaining an expected rate of growth over time, then that child is benefiting from the instructional program.

The purpose of this study was to extend findings on the effectiveness of bilingual education (e.g., Slavin & Madden, 1999a) on reading performance by further examining the effects of a schoolwide reading program on reading performance for students who received a 2-year ESL or maintenance bilingual program. The specific aim of this study was to compare the effects of a 2-year program that initially teaches a child to read in their native language to a program that primarily teaches the child to read in English, given an effective instruction program to determine which program has a greater positive effect on English reading ability when tested in English. Moreover, reading performance for the native Spanish-speaking students will be compared to native English-speaking students to evaluate differences between level and learning rates during third grade. The research questions were as follows.

Question #1. With the implementation of a schoolwide effective teaching reading program, is there a significant difference in reading performance between English-speaking students, bilingual students assigned to ESL only, or bilingual students initially instructed in their native language by the time students reach third grade and at four 8-week periods during third grade?

There are two a priori research hypotheses. First, based on prior research showing the effectiveness of SFA programs on reading (Slavin & Madden, 1999a, 1999b), participation in the SFA program should improve reading performance over time for all students. Second, controlling for SES and initial English language proficiency, the EPT maintenance program should consistently show greater improvement over time than the SFA program with ESL support when the EPT program is terminated (Greene, 1998; Thomas & Collier 1997, 2002).

Question #2: What is the rate of growth for each of these groups over a 1-year period of time in third grade given the initial second grade program assignment? In other words, to what extent does a schoolwide reading SFA program result in differences in the slope of achievement of reading performance between students who are English-speaking, bilingual students who were assigned ESL only, or bilingual students initially instructed in their native language?

Based on prior research that has demonstrated that all bilingual programs are effective (Willig, 1985), it is predicted that there will be differences in the level of reading performance between the native Spanish-speaking and English-speaking students, with English-speaking students obtaining superior levels. Alternatively, given the effectiveness on English-speaking students with programs incorporating effective instructional strategies (Greenwood et al., 1993), it is predicted Spanish-speaking students will most likely learn at rates equal to or greater than English-speaking students.

CHAPTER III

METHOD

In this chapter, the methods used to answer the research questions will be presented. Specifically, the experimental design used, details regarding the participants, the measures used for this study, and the procedures used for data collection are presented.

Experimental Design

This study used a quasiexperimental design using archival records to evaluate the effect of a teaching reading program on reading performance for native English- and for Spanish-speaking students receiving two different bilingual supplemental programs. This design was selected because participants were not randomly assigned to the three education programs examined in this study. The three educational program conditions examined in his study include: SFA program for English-speaking children (SFA only), SFA plus 45 minutes of ESL a day for bilingual Spanish-speaking children (SFA+ESL), and EPT, a version of SFA that teaches Spanish reading for bilingual Spanish-speaking children (EPT). See Appendix A for a graphic representation. The data was collected from a preexisting data set collected by the schools in which each student had a baseline reading score for third grade, and then four reading assessments during the school year.

Site

The setting was three elementary schools in Ogden City School District located in an urban area of northern Utah. Schools were included if the school had implemented

SFA during years 1998 - 2002, had similar ethnicity and SES demographic characteristics, and implemented one or both bilingual programs to be examined in this study (EPT and ESL). Specific school characteristics are presented in Table 1. A large percentage of students are from low SES families, which is reflected by the number of students who are eligible for free or reduced fee lunch. The district average is 49%, while all three elementary schools included in this study have a free or reduced lunch rate of 99%. The district ethnicity population consists of approximately 56% Caucasian, 36% Hispanic, 3% African-American, 1.5% Asian American, 1.5% American Indian, and .5% Pacific Islander (Ogden City School District, 2002). See Table 1 for each individual school's ethnic makeup.

Table 1

School Site Demographic Information

Demographics	School 1 %	School 2 %	School 3 %
SES			
Regular lunch	1	1	1
Rediced lunch	99	99	99
Ethnicity			
Caucasian	21	15	30
Hispanic	72	78	62
African American	4	37	5
LEP			
%	47	52	37
Service options	EPT, SFA+ESL	EPT, SFA+ESL	SFA+ESL

Limited English Proficient (ELL) Classification Process At Each Site

As mandated by law, each school must identify students who have limited English proficiency. Students were first identified in the participating schools from information obtained from a home language survey. Upon a child's initial enrollment for school, parents were asked via a written format or by an interpreter if there was another language spoken in the home other than English. If the parent indicated that another language was spoken at home, then the student was tested for English-language proficiency. The test results yielded categories of non-English proficient, limited-English proficient, and fluent-English proficient. If a student fell in the limited-English or non-English proficient categories, then he/she received alternative language services from the school as designated by the school's alternative language services team. The extent of the services available at each school varied. Two of the participating schools had Spanish instruction and ESL as service options, while the third school had only the ESL option available.

Once the child was participating in the alternative language program, a student's progress in academic achievement and English proficiency continue to be monitored. Students were transitioned out of alternative language services in one of two ways. First, a teacher could refer a student to the alternative language team for reclassification when a teacher felt a child had made substantial gains in English proficiency at any time during the school year. If referred, the student was retested on his/her language proficiency while the teacher gathered classroom work that documented student growth before a language team met to review the data. Based on progress data, the team decided to

maintain current services for the student, reduced services (i.e., transfer the student from primarily Spanish instruction to primarily English instruction with ESL support), or eliminated alternative language services with a monitoring phase of two years to make sure the child maintained adequate academic progress in the classroom. Reclassification also occurred if there was a change in the annual language proficiency testing and the alternative language services team decided to alter services.

Participants

In total, 121 third-grade students were identified and served as subjects in this the study. Third-grade students attending one of the participating schools in 2000-01 and 2001-02 were selected since this was the identified time in these schools when students who have been learning the general education curriculum primarily in Spanish were transitioned to classrooms that taught the curriculum in English. Subjects in this study included both English-speaking and bilingual Spanish-speaking students. All students who met the following inclusion criteria were included in the study.

First, all third-grade students were considered for the Spanish-speaking group if the schools identified the child as bilingual and the child's native language was Spanish as previously described in the ELL classifications procedures section. The Spanish bilingual students were then subdivided into two groups according to differences in educational programs. One group of students ($N = 24$) attended the ESL program since the beginning of first grade in addition to participation in SFA. The second group ($N = 25$) included students who were initially instructed in first grade to read the Spanish language through EPT. Students initially identified in first grade were included even if

students exited the program before third grade, because the objective was to examine the effects of the program, including its exiting criteria, at the time these students progress through third grade where all students were given reading lessons in English.

A third group of students ($N = 72$) included in this study were only receiving the SFA reading program. These students were included in the study if attended SFA since first grade, spoke English, were not identified at any time by the school as an ELL student, and did not receive alternative language services.

Subjects were selected via a three-step process. First, after obtaining approval for experimental procedures from the Utah State University human subjects review board, written permission was obtained from the principals of the three schools to collect data for this project. Second, students were required to attend the same school for grades 1-3, have baseline assessment data from the prior year, and participate in all STAR 8-week assessments their third-grade year. Third, students were then divided into one of the following three groups: English-speaking only students in SFA, bilingual Spanish-speaking students who attended ESL for part of the school day while in SFA, and bilingual Spanish-speaking students who were taught to read the Spanish language through EPT. Students from all three groups who met these criteria were included in the study. See Table 2 for participation demographic information.

Measures

English Oral Language Proficiency

Instrument. Oral language proficiency in "English was meased by the IDEA Proficiency Test (IPT) Oral Language Test (Del Vecchio & Guerrero, 1995). This is the

Table 2

Participant Demographic Information

Demo-graphics	SFA	ESL	EPT	SFA	ESL	EPT	SFA	ESL	EPT
Type of subjects									
N	30	6	19	15	14	6	27	4	0
Age									
M	9.32	9.36	9.28	9.42	9.39	9.40	9.41	9.29	--
SD	.34	.49	.28	.41	.32	.29	.37	.37	--
Retained									
N	1	0	0	1	0	2	1	0	00

in the process of classifying children as limited English proficient (see above). The IPT is administered at least annually to all students whose primary language is other than English. As part of the test administration, students are asked to do simple tasks, answer simple questions, and identify action verbs and nouns on stimulus cards (see Appendix B).

Normative data for this instrument were collected from a geographically diverse sample, with over 50% of this sample consisting of children from a Spanish-speaking background (Del Vecchio & Guerrero, 1995). The reported internal consistency is .99 for the Oral Language test. Concurrent validity was estimated by comparing test scores with teacher ratings of English language proficiency, which correlated at around .7. In addition, concurrent validity was shown with a correlation of .86 between the Language Assessment Scales, Language Rating Scales, and the Woodcock Language Proficiency Battery--Revised (Schrank, Fletcher, & Alavardo, 1996).

Administration and scoring. The IPT Oral Language is an individually administered oral assessment of English skills. Length of time to administer will vary depending on the student's fluency level. The average length of time is 14 minutes, with a range of 5 - 20 minutes (Del Vecchio & Guerrero, 1995). Tests are scored by calculating the total number of correct and incorrect responses and converting them into one of three categories: non-English-speaking (NES), limited English-speaking (LES), and fluent English-speaking (FES).

Reading

Instrument. Reading performance was obtained using the STAR Reading test that was administered during each subject's third-grade year in either the 2000-01 or 2001-02 school year. The STAR Reading test was specifically designed to evaluate a student's reading level within the Accelerated Reading program and is based on reading material presented in book lists generated by Renaissance Learning (Advantage Learning Systems, 2000). These lists include popular and traditional reading materials that have been evaluated by the Renaissance Learning Program to determine the reading grade level of each book. These books are then coded with colored tape that corresponds to specific reading levels to help teachers and students identify books that are at the child's current reading level. Students can then choose to read a book that matches the reading level that is determined by the STAR reading test and are used during a student's independent reading time. Students are also directly taught reading skills using the reading materials designated by the school for classroom instruction.

STAR Reading is a computerized, group-administered reading test that focuses on reading vocabulary to assess overall reading ability (Advantage Learning Systems, 2000). The test consists of multiple-choice questions asking the student to select words to fill blanks in a presented sentence that would best complete the sentence. The sentences are based on vocabulary taken from a specific book they have recently read as part of their reading program. Each test presents one sentence at a time with one missing word that is presented as a blank line. Below each sentence four words are presented in multiple-choice format. The student fills in a blank word by selecting the appropriate letter next to word that best completes the sentence from the list of words that are presented below. Items become easier or more difficult, depending on whether or not the student answers the question correctly. That is, if a student answers a test item correctly, a slightly more difficult question is then presented on the computer. However, if the student missed the test item, an easier sentence item is presented. The student is presented test items until he or she misses a certain number. The actual number of questions administered depends on the number of questions answered correctly and the pattern of correct and incorrect responses.

One major advantage to the STAR Reading test design is the ability to frequently monitor student progress during a school year. In order to decrease practice effects, students are administered different items each time a test is taken. Students are also tested on material from different books during each assessment session, thus pulling new questions based on a different book.

The STAR test yields both criterion-referenced and norm-referenced scores. Norm-referenced scores are based on a nationally representative sample of approximately

30,000 students in 47 states. The norm sample was stratified on three variables: geographic region, size of school district, and socioeconomic status. The scores for each student are derived from a computer-scoring program and class lists can be printed out and reviewed by teachers. Norm-referenced normal curve equivalent (NCE) scores were used for this study.

Studies supporting the validity and reliability of the test scores obtained with various groups of examinees from the STAR Reading Test are reported by test authors in their bulletin on reliability and validity (Advantage Learning Systems, 2000). Test-retest reliability and alternate form reliability for the STAR Reading for all grades is .94 and .95, respectively. Test-retest reliability for third grade is .87 and alternate form reliability is .86. Concurrent validity with 10 "high-stakes" achievement tests was established with correlations from .44 - .85, with most falling between .7 - .8 (see Appendix C).

Administration and scoring. The STAR Reading test was administered to each student five times. The first administration was conducted in May during a student's second grade year or the first two weeks of the third-grade year for students who did not have a score from May of the second grade year. The second administration was conducted at the end of October of a student's third-grade school year. Every 8 weeks thereafter during the school year a reading test was administered. That is, the third, fourth, and fifth administration was respectively given in January, March, and the end of April.

At this time, teachers brought their entire class to the school's computer lab classroom during a regularly scheduled weekly computer time. Before administering the test, teachers directed students to open the test program and to complete the test.

Although this test is administered via a computer, only very basic computer skills are needed. Additionally, students were previously taught required computer skills in weekly computer practice that had been conducted since students were in first grade. During the test administration, after students had opened the test program on their computers, teachers briefly instructed students to read each sentence and select the word that is missing by typing in the multiple-choice letter for the correct word shown on the screen. Although there is no time limit given for this test, the test manual reports that the test will take 10 minutes or less (Advantage Learning Systems, 2000). The student is required to complete the test independently and is given no additional help by the teacher while completing the test.

The computer program is designed to calculate the scores based on the students' responses and save the score for each student in a data base file set up for each classroom. The media specialist gives the teacher a printout for the entire class, which is generated by the program after each session. The teacher then sends the results to the SFA coordinator, who organizes the data in a schoolwide spreadsheet for data analysis. This process occurs every 8 weeks in the participating schools.

Dependent Variable

The primary dependent variable used for analysis was reading performance on the STAR Reading measure. A secondary analysis utilized the slope of growth trajectory from the STAR Reading. Normal curve equivalent (NCE) scores from the STAR Reading tests were used for data analysis. The NCE scores were calculated from the results of the STAR Reading tests from the baseline of the previous year and every 8 weeks thereafter

throughout the 9-month school year. NCE scores have a mean of 50 and a standard deviation of 21.06 (Gall, Borg, & Gall, 1996).

Data Collection Procedures

Data were collected in the following manner. First, students who would have been in third grade for either the 2000-01 and 2001-02 school year were identified. The cumulative file for each student was then examined to see if the student matched the selection criteria of attending their respective school since first grade. If the student matched these criteria, their school ID number was recorded on a data collection sheet (see Appendix D). Also recorded on this sheet was the number of years attended at that school, birth month and year (to guard confidentiality), whether the student had been retained, bilingual classification, original language proficiency score if bilingual (i.e., non-English proficient, limited-English proficient, fluent English proficient), and, if bilingual, whether they were initially placed in EPT or SFA with ESL.

This list was then taken to the archival database of the computer program for the STAR Reading assessments. A list of names and ID numbers of students who would have been in third grade in either 2000-01 or 2001-02 was obtained. This list was kept separate from the data collection sheet, never left school property, and was solely used to search for students in the STAR Reading computer database. This method was utilized because a student ID search in the STAR Reading database was not always possible. A printout of the student's test score history was obtained, the printout was matched up with the corresponding student ID number on the data collection sheet, and the student's name was marked out until unidentifiable on the STAR Reading printout.

The printouts contain the STAR Reading scores that were generated for the teacher every 8 weeks during that school year. In order to be included in the study, students needed to have a baseline data point and participated in each 8-week assessment. Data were not collected at School 3 for the 2000-01 school year due to a computer error, which resulted in a lack of computer-accessible data, and no paper hard copy of the same data being available.

Independent Conditions and Procedures

Three educational program conditions were compared in this study including: SFA program for English-speaking children (SFA only), SFA plus 45 minutes of ESL a day for bilingual Spanish-speaking children (SFA+ESL), and Éxito Para Todos (EPT), a version of SFA that teaches Spanish reading for bilingual Spanish-speaking children (EPT). A brief description and program procedures for each of the three programs are presented in the following three sections.

Success for All (SFA) program. In this condition, students participated in a schoolwide reading program, SFA, for 2 school years (i.e., first and second grade) previous to this study. Students continued to participate in the SFA program during third grade (i.e., the year data was taken for this evaluation).

Although formal documentation of program fidelity was not conducted in this study, there are built-in components to the SFA training model used to ensure consistent initial implementation as well as on-going fidelity to the program. Teachers and staff are first introduced to the SFA program during a three-day workshop conducted by the national organization. In addition, the SFA coordinator from the school and the principal

attend a five-day "New Leaders Conference." Next, local SFA coordinators use procedural checklists to evaluate the degree of implementation by conducting classroom observations. SFA coordinators observe implementation in teacher classrooms for three days after training sessions have been completed. Based on the information obtained from these observations, the local SFA coordinator refines or retrains teachers on any missing program components.

Once the initial training and classroom observations are conducted, each school has a full-time coordinator whose primary responsibility is to monitor the progress and implementation of the program. The coordinator continues to conduct classroom visits, coaching, and team meetings. Moreover, teachers are encouraged to develop coaching partnerships, help maintain implementation of the program and problem-solve any difficulties. Implementation visits are made by the national organization at least 3 - 4 times per year initially and then decrease depending on the needs of the school and students. Training is also conducted by the national organization for the local schools at the beginning of each school year. The exact number of visits conducted at each school was not available. The SFA coordinators for School 1 and 2 were the same coordinators from the inception of the program, while School 3 had a change in SFA coordinator at the beginning of the 2001-02 school year. All coordinators reported adhering to the program overall, using annual training to introduce new teachers to the program and review program components for continuing teachers, and working with individual teachers when program adherence became an issue.

Success for All (SFA) and ESL program. In this condition, students participated in a pullout ESL program in addition to the schoolwide SFA 90-minute reading program.

Participation in these two programs was conducted during the first and second grade prior to this study and continued during the third-grade year in which data was collected. The pullout ESL program consisted of instruction for 45 minutes a day for ELL students. At this time, students were taught using the curriculum "Into English," which is a theme-based curriculum using songs, poems, and stories to help facilitate the acquisition of English. None of the curriculum content uses the student's native language (Tinajero & Schifini, 1997). Teachers will occasionally use Spanish to give directions (e.g., "Sit down," "Look at this") at the beginning of the year for students recently emigrated from Spanish-speaking nations. Pantomime is used with more limited English-speaking students for instruction.

An example illustrates how this would work in a classroom with students at different English proficiency levels. If the theme for the unit is a beaver, the teacher might point to a beaver and have the student name the animal. The next level would be to ask a student where the beaver is in the (book, room, story). Or another strategy would be to do an oral "fill in the blank" of "An animal that builds dams is a _____." Still a higher level of skills would have the students tell the teacher about the beaver.

Éxito Para Todos (EPT) SFA program. In this condition, students participated in EPT, the SFA program for teaching students to read the Spanish language, which is initiated in first grade. Native Spanish-speaking students were placed in the program if designated by the alternative language services team for the school (see ELL classification procedures for more information). The EPT curriculum approach is the same as the English SFA reading curriculum approach and employs the same effective teaching strategies, but uses those same strategies to teach students to read the Spanish

language. Once students completed the portion of EPT that covers basic reading skills and obtained a mastery criterion in the curriculum, they moved into an SFA classroom where they began to be taught how to read in English with ESL support.

CHAPTER IV

RESULTS

Preliminary Analyses

Descriptive statistics for student per experimental condition, SFA, ESL+SFA, and EPT, are given in Table 3. In quasiexperimental designs, it is essential to examine comparability of samples. Thus, the reading groups were first compared on demographic variables of age (for the English, SFA+ESL and EPT groups) and initial IPT score (for the SFA+ESL and EPT groups) using a one-way ANOVA and chi-square analysis, respectively. No significant differences between the reading programs were found for age, $F(2, 114) = .399$, $p = .672$, or initial IPT, $\chi^2 (1, 49) = 1.647$, $p = .199$.

Reading scores were examined between the three schools to see if any differences existed between the three schools that might suggest uneven program implementation between schools. To examine differences between reading programs in the three schools, a repeated measures mixed model ANOVA was conducted to determine if any significant differences on reading performance with school membership (i.e., schools 1, 2, 3) as the between subject factor and testing time (i.e., administrations 1 through 5) as the with-in subjects or repeated factor. Because there were differences between the schools in terms of type of bilingual program, only the English-speaking group from each school was used for the comparison. The significance level used for this and all other analyses of statistical significance is $p <. 05$. Table 4 shows the means, standard deviations, and ranges of the English-speaking groups. Results of this analysis reveals no significant difference between schools, $F(2, 69) = .174$, $p = .840$, no significant difference over

Table 3

Means, Standard Deviations, and Ranges of Age and Counts with Percentages for Initial IPT Scores for SFA, ESL+SFA, and EPT Groups

Characteristic	SFA	ESL+SFA	EPT
Age			
M	9.41	9.29	9.41
SD	.369	.375	.263
Range	8.92 - 10.33	9.00 - 9.83	9.08 - 9.75
Initial IPT score			
LES (%)	--	14 (53.3%)	10 (40%)
NES (%)	--	10 (41.7%)	15 (60%)

Table 4

Means, Standard Deviations, and Ranges of Star Reading Scores for English-Speaking Students in the Three Participating Schools over the Five Assessment Periods

Characteristic	Baseline	1st assessment	2nd assessment	3rd assessment	4th assessment
School 1 (*n* = 30)					
M	29.65	30.53	32.53	32.99	34.99
SD	23.57	21.08	21.30	21.34	19.69
Range	1.0 - 83.6	1.0 - 77.2	1.0 - 83.6	1.0 - 77.2	1.0 - 73.4
School 2 (*n* = 15)					
M	34.05	34.55	37.35	34.34	35.87
SD	21.48	23.80	24.19	17.98	18.83
Range	1.0 - 82.7	1.0 - 86.9	6.7 - 86.9	10.4 - 79.6	6.7 - 79.6
School 3 (*n* = 27)					
M	32.41	33.43	35.11	37.31	35.38
SD	21.59	19.79	22.70	19.59	19.47
Range	1.0 - 79.6	1.0 - 63.5	1.0 - 77.0	1.0 - 68.5	1.0 - 71.8

time, $F(4, 276) = 2.139$, $p = .076$, and no significant interaction, $F(8, 276) = .486$, $p = .866$. It appears there are no differences on reading scores of English-speaking students between the three schools during all five assessment periods, suggesting that implementation of the reading program did not differ across schools.

A second repeated measures ANOVA was conducted with school year (2000-2001, 2001-2002) as the between-subject factor and time of testing (administrations 1 - 5) as the within-subject factor to estimate if possible fluctuations between years (changes in school personnel, additional experience with the SFA program, etc.), lead to significant differences in reading scores. All three groups of students (i.e., English-speaking, ESL, EPT) were used in this analysis. Because the sphericity assumption was not met, a Greenhouse-Geisser correction was employed. Means, standard deviations, and ranges are shown in Table 5. Results indicated a significant main effect of time $F(3.648, 434.066) = 6.577$, $p < .001$, but a nonsignificant main effect of school year, $F(1,119) = 2.214$, $p = .139$, and a nonsignificant interaction, $F(3.648, 434.066) = .525$, $p = .701$. In summary, these results reveal that scores improve over time during both school years, but are not affected by the year during which data were collected.

Primary Analyses

Tests of Statistical Significance

Means, standard deviations, and range of scores for the reading programs on reading performance at each time of testing are presented in Table 6. A repeated measure ANOVA was conducted to analyze the main effects and interaction between the between-subjects reading group conditions (SFA, SFA+ESL, or EPT) and repeated

Table 5

Means, Standard Deviations, and Ranges of STAR Reading Scores for All Participating Students Each School Year over the Five Assessment Periods

Year	Baseline	1st assessment	2nd assessment	3rd assessment	4th assessment
2000-2001 (*n* = 36)					
M	22.07	22.55	26.01	26.49	26.41
SD	18.88	16.92	19.41	17.63	17.41
Range	1.0 - 70.1	1.0 - 60.4	1.0 - 60.4	1.0 - 65.6	1.0 - 69.3
2001-2002 (*n* = 85)					
M	27.73	29.61	30.28	31.30	32.54
SD	22.06	20.48	21.68	20.62	19.36
Range	1.0 - 84.6	1.0 - 86.9	1.0 - 86.9	1.0 - 79.6	1.0 - 79.6

Table 6

Means, Standard Deviations, and Ranges of STAR Reading Scores for SFA, ESL+SFA, and EPT Groups over the Five Assessment Periods

Program	Baseline	1st assessment	2nd assessment	3rd assessment	4th assessment
SFA (*n* = 72)					
M	31.60	32.46	34.50	34.89	35.32
SD	22.17	20.96	22.2	19.85	19.17
Range	1.0 - 84.6	1.0 - 86.9	1.0 - 86.9	1.0 - 79.6	1.0 - 79.6
ESL (*n* = 24)					
M	17.28	20.85	20.62	21.69	21.85
SD	15.44	13.14	15.55	15.80	16.16
Range	1.0 - 61.0	1.0 - 52.1	1.0 - 63.5	1.0 - 76.0	1.0 - 77.0
EPT (*n* = 25)					
M	18.48	19.64	21.24	23.26	25.99
SD	18.57	17.03	17.32	19.30	17.07
Range	1.0 - 68.5	1.0 - 62.9	1.0 - 65.6	1.0 - 62.9	1.0 - 65.6

measure time (administrations 1 - 5). Because the sphericity assumption was not met (Greene-Geisser $\varepsilon = .92$, Huynh-Feldt $\varepsilon = .97$), a Geisser-Greenhouse correction was used to offset the bias generated by the failure to meet the sphericity assumption, although $\varepsilon > .90$ is generally considered a small departure from sphericity (Grimm & Yarnold, 2000).

Effect sizes were also calculated to provide additional evaluation of the magnitude of the effects of the reading program for each group over time. Partial eta squared (η_p^2) was used due to the presence of more than two groups with a repeated measures design. Partial eta squared gives the proportion of the variance each factor contributes to the overall sample (Cohen, 1977). This would be comparable to the interpretation of R^2 when considering the magnitude of effect of a correlation coefficient. Cohen gave a reference for the magnitude of effect as .01 for a small effect, .059 for a medium effect, and .138 for a large effect.

The results of this analysis indicate a significant main effect for reading program, $F(2, 118) = 7.166$, $p = .001$, with a medium magnitude of effect ($\eta_p^2 = .108$). Time is also a statistically significant factor, $F(3.651, 430.825) = 6.774$, $p < .001$, with small magnitude of effect ($\eta_p^2 = .054$). However, no significant interaction effect existed, $F(7.338, 432.929) = .637$, $p = .732$, $\eta_p^2 = .011$) indicating the passage of time did not have a differential effect on instruction on any group of students. A graphical representation of reading score differences between SFA, ESL+SFA, and EPT reading groups is shown in Figure 1.

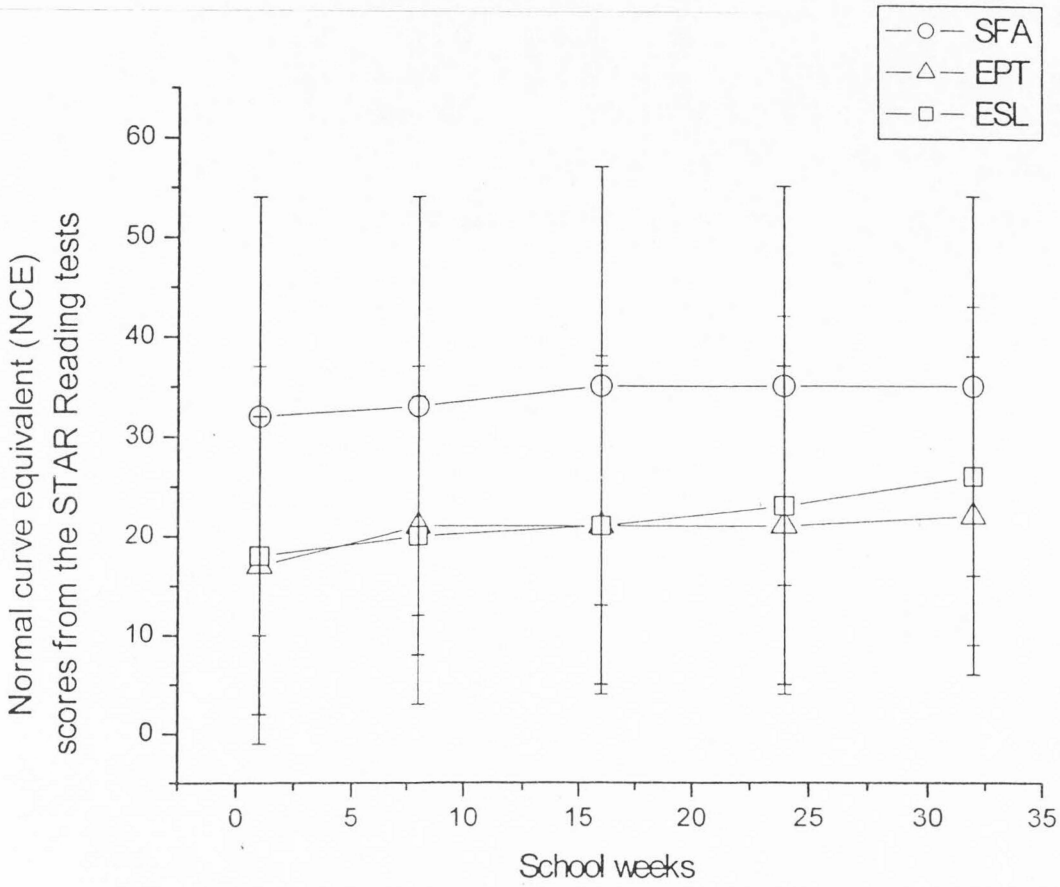


Figure 1. Mean NCE scores on the STAR reading test for SFA, SFA+ESL, and EPT groups at five test administration times.

Because a statistically significant main effect finding was found for the reading program, post hoc analysis using Tukey's HSD was employed to further analyze significant differences between the three reading groups at each test administration time. Follow-up tests showed that the SFA group was statistically significantly higher at the $p < .05$ significance level than EPT as well as the SFA+ESL groups at baseline, the 1st

assessment, 2nd assessment, and 3rd assessment. At the 4th assessment, SFA is significantly higher than SFA+ESL at the $p < .05$ significance level; however, there is no significant difference between SFA and EPT at the $p < .05$ significance level. There was no statistically significant difference between EPT and SFA+ESL at any of the assessment periods.

A polynomial contrast corresponding to linear, quadratic, cubic, and quartic effects was conducted to determine how reading scores changed over time and if the steepness of slope lines for each reading group significantly differed from one another. There was a significant main effect for time on the linear component, $F(1, 118) = 19.28$, $p < .001$, $\eta_p^2 = .14$, but no significant interaction effects for all polynomial terms. Thus, all group reading scores linearly increased with similar steepness of slope.

With time as a significant factor, an additional analysis was conducted to see the relative effect of time for each group. A mean effect size, using Cohen's *d*, was conducted for each group at each assessment point during the school year compared to baseline. For example, the mean score of 32.46 for SFA at the 1st assessment was compared to the baseline mean of 31.60 (as noted on Table 6). Table 7 shows the effect sizes for the three groups. A rule of thumb for magnitude of effect is .20 for a small effect and .40 for a medium effect (Kotrlik & Williams, 2003).

Slope Analysis

An additional analysis conducted in this study examined the effect of reading programs on the slope of achievement or the rate of growth inindividual student performance from baseline to the end of one school year. Appendix E depicts a sample

Table 7

Mean Effect Sizes (Cohen's D) for SFA, ESL, and EPT Groups at Four Assessment Points Compared to Baseline Performance

Program	1st Assessment	2nd Assessment	3rd Assessment	4th Assessment
SFA (*n* = 72) Cohen's *d*	.04	.13	.16	.18
ESL (*n* = 24) Cohen's *d*	.25	.22	.29	.30
EPT (*n* = 25) Cohen's *d*	.07	.16	.26	.43

of a growth rate chart from one of the students in the study. This figure models student learning by showing the rate at which a student is acquiring knowledge. Each data point represents a NCE score obtained at baseline and each of the 8-week assessments. The slope of the reading growth trajectory line represents the overall trend of a student's reading ability. The solid line is a regression line fit to each student's scores using the ordinary least squares method (Deno, Fuchs, Martson, & Shin, 2001; Good & Shinn, 1990). The resulting beta values then represent the slope of that line and were used for further analysis.

Table 8 shows the descriptive statistics for the individual slope based on the beta values. A one-way ANOVA was then conducted on the variable of reading program using these beta values. This analysis revealed no significant difference between the reading programs related to slope, $F(2, 118) = 1.211$, $p = .301$, $\eta_p^2 = .020$.

Table 8

Means, Standard Deviations, and Ranges of Beta Values for SFA, ESL+SFA, and EPT Groups

Statistics	SFA	ESL+SFA	EPT
Beta Values			
M	.9511	.7383	1.8644
SD	2.8724	2.1625	3.2193
Range	-5.43 - 8.10	-2.54 - 5.40	-5.87 - 7.90

CHAPTER V

DISCUSSION

With the rapid increase in ELL populations, schools are increasingly confronted with the double demand of teaching reading skills within a critical developmental period to students who are also learning English. Thus, one major issue in bilingual research is whether the quality of instruction is as important as or more important than the language of instruction during early reading instruction. The primary purpose of this study was to evaluate differences between ELL students who were given reading instruction in either Spanish or English while receiving a schoolwide program, SFA, that employed quality instruction.

Results of this study first indicated that ELL students were performing at a significantly lower level of reading performance than native English-speaking students at the beginning of third grade regardless of the language of the initial reading program given in first and second grade. Further, the results of this study revealed no significant difference in reading performance between the ELL students receiving instruction in Spanish (EPT) and or in English with ESL at the initial third-grade assessment before all students received instruction in English. These results are surprising given the briefer period of time that EPT students were reading in English.

As students progressed in the SFA program with English reading instruction, the achievement gap between native English speakers and native Spanish speakers noted at the beginning at third grade persisted for 24 weeks. By the fifth assessment (32 weeks), post hoc analysis and the graphical representation of the means show that the gap in

performance between the EPT group and the SFA group significantly narrowed while the SFA+ESL only group's score remained significantly lower than SFA. However, there was no significant difference between the two Spanish-speaking reading groups.

Because this study employed a quasiexperimental design, some other unmeasured variable may influence these results. However, because students came from similar low SES backgrounds and had similar English language proficiency scores at entry into a bilingual program, these factors did not differentially affect the pattern of results between the two Spanish-speaking groups in this study. This is important because of the variability between language experiences amongst ELL students is a common confound in many previous studies. Moreover, English-speaking and Spanish-speaking students consisting of low SES background is highly correlated with low reading results regardless of language experience (Donahue, Voelkl, Campbell, & Mazzeo; 1999).

When the mean effect sizes were shown for each group over time, an interesting pattern emerged. The changes over time for the SFA group were marginal; however, the two ELL groups showed a greater magnitude of change in progress over time, obtaining a medium magnitude of effect when comparing the baseline and final assessment of the school year.

Given previous research findings (Greene, 1998; Thomas & Collier, 1997, 2002; Willig, 1985), it was expected a priori that instructing bilingual students to read in their native language first would provide an added benefit for students in their reading scores. Results in this study indicated a trend at the end of third grade where the EPT student reading scores started improving over the SFA+ESL group. The effect size differences between groups was minimal, at best, and showed that type of reading program was not a

significant contributor to the variance among reading scores. These results are similar to the findings of Ramirez (1992) who also found no difference between immersion and transitional approaches although initial English language proficiency was controlled for and an intensive program such as SFA was utilized in this study. In addition, the lower effects of native language instruction on reading performance could be related to students still acquiring cognitive-academic language proficiency (CALP; Cummins, 1999; Hakuta et al., 2000). Thus, differential outcomes based on the amount of native language instruction may not be expected to be seen at this early stage in a student's career (Ramirez; Thomas & Collier) and additional time may be necessary to acquire more advanced skills than are measured by language proficiency tests, such as vocabulary level or background knowledge in English. An indication of this could be the larger effect size for EPT at the end of third grade. However, this finding may be further evidence of the lack of substantial benefit of native language instruction over English-only methods (Rossell, 1998; Rossell & Baker, 1996).

This finding is also consistent with the studies of EPT conducted in Philadelphia and California for English language proficiency (Slavin & Madden, 1999a). However, methods used in this study extended these findings by using multiple assessments across a school year instead of the sole reliance on a pre- and posttest. Also, this study continued to track students after they had been transitioned out of bilingual services and included an English-only control or comparison group, which was advanced by Greene (1998) as an important consideration in evaluating the methodological soundness of studies of bilingual education. The comparison of the ELL groups to the English-only control clearly demonstrated significant positive gains in reading scores for all reading

groups during third grade that was maintained over time (32 weeks). However, the magnitude of the effect sizes would indicate these gains were most pronounced for the EPT and SFA+ESL groups. Importantly, the overall statistical trend based on slope analysis suggested that ELL students on average were learning to read in the program at rates similar or higher than native English-speaking students when slope was determined across 32 weeks of instruction.

It is noteworthy that the variability in the data demonstrated by large ranges and large standard deviations suggested large individual differences in student performance within each reading program. But if percentages of students whose scores reflect a substantial performance deficit that would suggest a need for more intensive instruction are considered, there were approximately 40% (range 38 - 42%) of students whose growth trajectory (slope) did not increase in each of the three reading groups. Additionally, there were similar percentages of students whose scores fell 1 standard deviation below the mean of their group mean performance in May (range 9 -16%).

Limitations of Research and Directions for Future Research

There are several limitations related to the outcomes of this study that can be addressed in future studies. First, the data collected in this study were collected using archival records. Thus, the experimenter was not able to have complete control of the fidelity of the implementation of reading programs or the initial data collection process. Anecdotally, the information obtained for training time for teachers and administrators, as well as time spent on program implementation by the local and national

representatives of SFA (and EPT) suggest a well-supported program. Teachers were provided with 3 days of initial training, one day of training a year as well as a full-time program coordinator who provided classroom assistance. Schoolwide scheduled reading time also ensured that all students consistently received 90 minutes of reading instructional in on-level reading groups. However, there was no independent verification conducted when the SFA program was implemented.

Secondly, the sample size was not very large in these groups. The effect size was calculated for each of these groups to attempt to offset this factor. However, there was a great deal of variation with the treatment groups, as demonstrated by the large standard deviations for each of the means reported. Thus, there appear to be no effects when, in fact, there could be treatment effects which are obscured by outliers measured in the study. With a small sample size, these outliers are difficult to detect, much less exclude from the already small sample size. Obtaining an adequate sample size is important to determine if there are outliers, which could obscure results. This is a difficult task given that this is a highly mobile population (Snow et al., 1998). In this investigation, only 28% of students had been at the same school at least 3 of 4 years. Extracting a larger sample for a similar study in the future would aid greatly in the replication of this study. If possible, finding a more stable Spanish-speaking population from which to draw a sample would help eliminate this possible threat to validity.

Thirdly, the STAR Reading Instrument was problematic in its measurement. A number of students ($N = 24$) across all three groups had a NCE score of 1.0 (lowest score possible) in at least three of the five assessment periods. This suggests an inadequate floor. As a result of this floor effect, it is unclear what reading skills these students had

mastered. In other words, were these students who were struggling to learn their letters or were they students who got one or two of the items correct? Additionally, it could not be determined if there were students who had higher reading abilities but were simply unmotivated to complete the test.

There were also many students with large variability in NCE scores. One individual's NCE score would be at 30, for example, then drop to 10, and then go to 35 the next assessment period. Due to procedures used in this study it was not possible to determine if variable results were due solely to reading skills acquired or whether other factors, such as motivation or distraction during testing sessions or interest in the book chosen, affected scores. A better measurement instrument would also be appropriate. The floor effects mentioned earlier may have skewed the results, as well as possible distractions during testing. An individually administered instrument may be preferable to minimize distractions. A measurement instrument with greater sensitivity, such as curriculum-based measurement (CBM) oral reading fluency rates, would show changes over time, both minute and grand.

When gathering data, the only demographic variables recorded were the student's age and whether they had been retained. Other background information, such as gender, free/reduced lunch status, and ethnicity could have been collected to see what effects these variables had on the results of the study. Without those data collected, it is unknown whether the sample was truly representative and whether these demographic variables played a role.

Another limitation is the generalizability of this study. It was conducted in one inner city school district in the Rocky Mountain region and is not necessarily

representative of what would occur in another geographic area or nationally. Further, this study was conducted for only the third grade. This may not reflect achievement or differences for younger or older grades. Definitive conclusions cannot ever be made based on a single study and must be considered in light of other research conducted in this area. For example, some studies have shown a greater disparity between students' reading achievement over the course of their education, depending on the amount of language instruction received (Ramirez, 1992; Thomas & Collier, 1997, 2002). A longitudinal study, such as the ones conducted by Ramirez or Thomas and Collier, which incorporates more of a student's educational career and utilizes the type of dynamic assessment techniques used in this study, would help to draw more definitive conclusions on the growth trajectories of these populations. Another extension of this research would be to conduct it at each grade level (fourth, fifth, etc.) to see if a similar pattern emerges or if there are differences due to grade. Higher grades may also be less sensitive to the floor effects mentioned previously.

Longitudinal studies in bilingual education have shown gains in reading over time, especially with a greater amount of native language instruction (Ramirez, 1992; Thomas & Collier, 1997, 2002). This study did only take into account gains for one year, but if the trend found in this study continued into later grades, reading levels of students in the SFA group would approach the national average and reading levels of students in the SFA+ESL and EPT groups would be within 1 standard deviation of the national average by the end of their elementary school years. This would equip bilingual students with greater reading ability, an important factor in dropout rates (McMillan, Kaufman, & Klein, 1997). Moreover, it would defy the "Matthew effect" documented for students

with low reading scores in early grades and/or come from low SES backgrounds (Donahue et al., 1999; Shaywitz et al., 1995; Stanovich, 1986) in that these students are showing gains, not losses, in reading scores over time. This finding was, of course, only found for one grade level at three schools in one district. Further replication would be needed to confirm that this phenomenon is occurring.

The differences noted at the end of third grade for the gap between SFA and EPT, and SFA and SFA+ESL, plus the jump in effect size for EPT at the end of third grade, could be an indication of where the trajectories for native language instruction and ESL services diverge, as seen in other studies (Ramirez, 1992; Thomas & Collier, 1997, 2002). The high degree of variability along with the lack of data beyond third grade makes it impossible to ascertain whether this is the case, but it could be a possible area of investigation for future researchers.

Another possible extension of this study would be to conduct a follow-up study for this same group of students in 5 - 10 years, when these students are in junior high and high school, and evaluate their academic level. Thomas and Collier (1997, 2002) have shown a "sleeper effect" in which students show similar achievement at early grades, but then students show differential gains over time depending on the amount of native language instruction received. A follow-up study could test this hypothesis of Thomas and Collier and be another way to see the long term implications of native language instruction for these students. This study could also investigate whether the reading gains demonstrated in third grade lead to greater long term academic outcomes.

Implications for Practice

The type of methodology utilized for this study is a unique contribution to this body of literature. Few of the major bilingual education studies cited in this investigation (Greene, 1998; Ramirez, 1992; Rossell & Baker, 1996; Thomas & Collier, 1997, 2002; Willig, 1985) used a methodology that examined ongoing results during a school year, as well as an analysis of the slope to find the growth rate of the student over time. Similarly, investigations of the effectiveness of SFA and EPT (Herman et al., 1999; Slavin & Madden, 1995, 1999a, 1999b) have employed a pre- postmodel looking at yearly changes without consideration for growth rate.

Using this methodology, the results have major implications for identification of students for additional services such as prereferral services or special education. Schools are challenged to meet two objectives. First, in order to be proactive, struggling students need to be identified early for intervention services. For ELL students, early identification must occur while students are still learning English since it may take 2 to 7 years for students to become proficient (Cummins, 1999). Yet, differentiating between an ELL student who is struggling due to language from other learning problems must be accomplished if identification procedures are to be nondiscriminatory. Second, schools are trying to diminish overrepresentation of minority students while avoiding underrepresentation of these students (Gersten & Woodward, 1995).

Initially, the results in this study revealed that performance levels at the first four administrations suggested that more ELL students were performing at a lower level than English-speaking students regardless of early reading program, resulting in more students

who may be identified for intense services if language influences are not considered. Alternatively, when looking at growth rates of students, the ELL students in both groups had a similar pattern of growth to the English-speaking group. Thus, in attempts to identify students at-risk, we can look for ELL students to have a similar growth pattern to other students, provided they have been given quality instruction for a few years. Those students who show much lower growth rates would, according to this study, be at-risk for failure and a natural target for intervention.

This study further supports that slope analysis may be a more equitable model of student learning (Deno et al., 2001) that would give practitioners a better idea of what "normal" is for students to achieve at a given point in a year. Then intervention can be implemented sooner for students who are not successful at a given point in the school year, rather than the next year when the end-of-year testing results are reviewed. Moreover, using growth rates would give practitioners a better idea that students just "need more time" to increase their English language proficiency and which need targeted special interventions for other learning difficulties.

In this study, the amount of native language used in the program did not show much effect. However, having an effective program did show growth over time for all students. Perhaps the incorporation of effective instruction practices for a given program should be as important a consideration as the amount of native language versus English contained in the program.

REFERENCES

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Advantage Learning Systems. (2000). *STAR Reading: Understanding reliability and validity.* Wisconsin Rapids, WI: Renaissance Learning.

August, D., & Hakuta, K. (1997). *Improving schooling for language-minority children.* Washington, DC: National Academy Press.

Baker, S., & Hakuta, J. (1997). *Bilingual education and Latino civil rights.* Paper presented for the Harvard Civil Rights Project. Retrieved July 5, 2002, from http://www.law.harvard.edu/groups/civilrights/conferences/latino/papers/bilingual/bilingual.html

Calderón, M., Hertz-Lazarowitz, R., & Slavin, R. (1998). Effects of bilingual cooperative integrated reading and composition on students making transition from Spanish to English reading. *The Elementary School Journal, 99*(2), 153-165.

Carter, T. P., & Chatfield, M. L. (1986). Effective bilingual schools: Implications for policy and practice. *American Journal of Education, 95*, 200-232.

Castañeda vs. Pickard, 648 F.2D 989 (5th Cir.), (1981).

Cohen, J. (1977). *Statistical analyses for the behavioral sciences.* New York: Academic Press.

Council of Chief State School Officers. (1998). *State education indicators with the a focus on Title I.* Washington, DC: Author.

Cummins, J. (1999). *BICS and CALP: Clarifying the distinction.* (ERIC Document Reproduction Service No. ED 438551)

Del Vecchio, A., & Guerreo, M. (1995). *Handbook of English language proficiency tests.* Albuquerque, NM: Evaluation Assistance Center, Western Region.

Deno, S. L., Fuchs, L. S., Martson, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*(4), 507-524.

Donahue, P. L., Voelkl, K. E., Campbell, J. R., & Mazzeo, J. (1999). *NAEP 1998 reading report card for the nation and the states.* Washington, DC: National Center for Educational Statistics.

Fletcher, T. V., Bos, C. S., & Johnson, L. M. (1999). Accommodating English language learners with language and learning disabilities in bilingual education classrooms. *Learning Disabilities Research & Practice, 14*(2), 80-91.

Fuchs, L. S., & Fuchs, D. (1998). Treatment validity: A unifying concept for reconceptualizing the identification of learning disabilities. *Learning Disabilities Research & Practice, 13*(4), 204-219.

Gall, M. D., Borg, W. R., & Gall, J. P. (1995). *Educational research: An introduction* (6th ed.). New York: Longman.

Gersten, R., & Baker, S. (2000). What we know about effective instructions practices for English-language learners. *Exceptional Children, 66*(4), 454-470.

Gersten, R., & Woodward, J. (1994). The language-minority student and special education: Issues, trends, and paradoxes. *Exceptional Children, 60*(4), 310-322.

Gersten, R., & Woodward, J. (1995). A longitudinal study of transitional and immersion bilingual education programs in one district. *The Elementary School Journal, 95*(3), 223-239.

Goldenberg, C., & Gallimore, R. (1991). Local knowledge, research knowledge, and educational change: A case study of early Spanish reading improvement. *Educational Researcher, 20*(8), 2-14.

Good, R. H., & Shinn, M. R. (1990). Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*(2), 179-193.

Greene, J. (1998). A meta-analysis of the Rossell and Baker review of bilingual education. *Bilingual Research Journal, 21*(2-3), 103-22.

Greenwood, C. R. (1991). Longitudinal analysis of time, engagement, and achievement in at-risk versus non-risk students. *Exceptional Children, 57*(6), 521-535.

Greenwood, C. R. (1996). The case for performance-based instructional models. *School Psychology Quarterly, 11*(4), 283-96.

Greenwood, C. R., Delquadri, J., & Bulgren, J. (1993). Current challenges to behavioral technology in the reforming of schooling: Large-scale, high-quality implementation and sustained use of effective educational practices. *Education and Treatment of Children, 16*(4), 401-440.

Grimm, L. G., & Yarnold, P. R. (2000). *Reading and understanding more multivariate statistics.* Washington, DC: American Psychological Association.

Hakuta, J., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency?* Santa Barbara, CA: University of California Linguistic, Minority Research Institute.

Herman, R., Aladjam, D., McMahon, P., Masem, E., Mulligan, I., Smith, O., O'Malley, A., Quinones, S., Reeve, A., & Woodruff, D. (1999). *An educator's guide to schoolwide reform.* Arlington, VA: Educational Research Service.

Krashen, S. D. (1998). A note on Greene's "A meta-analysis of the effectiveness of bilingual education." Retrived March 23, 2002, from http:// ourworld.compuserve.com/homepages/ JWCRAWFORD/ Krashen2.htm

Kotrlik, J. W., & Williams, H. A. (2003). The incorporation of effect sizes in informational technology, learning, and performance research. *Informational Technology, Learning, and Performance Journal, 21*(1), 1-7.

Lam, T. C. (1992). Review of practices and problems in the evaluation of bilingual education. *Review of Educational Research, 62*(2), 181-203.

Lau vs. Nichols, 414 U. S. 563, (1974).

Lambert, W. E. (1992). Challenging established views on social research: The power and limitation of research. *American Psychologist, 47*(4), 533-42.

McMillen, M. M., Kaufman, P., & Klein, S. (1997). *Dropout rates in the United States: 1995.* Washington, DC: National Center for Educational Statistics. (ERIC Document Reproduction Service No. ED 410 370)

Mosteller, F., Light, R. J., & Sachs, J. A. (1996). Sustained inquiry in education: Lessons from skill grouping and class size. *Harvard Education Review, 66*(4), 797-842.

National Center for Educational Statistics. (1999). *NAEP 1998 reading: A report card for the nation and the states.* Washington, DC: Author.

National Center for Educational Statistics. (2000a). *NAEP summary data table.* Washington, DC: Author.

National Center for Educational Statistics. (2000b). *Dropout rates in the United States: 2000.* Washington, DC: Author.

Ochoa, S. H., & Perez, R. J. (1995). The Journal of Educational Issue of Language Minority Students, 15. Retrieved March 23, 2004, from http://www..ncbe.gwu.edu/miscpubs/ jeilms/vol15/appraising.htm

Ochoa, S. H., Rivera, B. D., & Powell, M. P. (1997). Factors used to comply with the exclusionary clause with bilingual and limited-English proficient pupils: Initial guidelines. *Learning Disabilities Research & Practice, 12*(3), 161-167.

Ogden City School District. (2002). *Ogden City Schools annual report.* Ogden, UT: Author.

Ramirez, J. D. (1992). Executive summary--Final report: Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual education programs for language-minority children. *Bilingual Research Journal, 16,* 1-62.

Rossell, C H. (1998). Mystery on the bilingual express: A critique of the Thomas & Collier study. Retrieved March 20, 2002, from http://www.ceousa.org/ READ/ collier.html

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

Schmidt, R. J., Rozendal, M. S., & Greenman, G. G. (2002). Reading instruction in the inclusion classroom: Research-based practices. *Remedial and Special Education, 23*(3), 130-140.

Schrank, F. A., Fletcher, T. V., & Alvardo, C. J. (1996). Comparative validity of three English oral language proficiency tests. *Bilingual Research Journal, 20*(1), 55-68.

Shaywitz, B. A., Holford, T. R., Holahan, J. M., Fletcher, J. M., Stuebing, K. K., Francis, D. J., & Shaywitz, S. E. (1995). A Matthew effect for IQ but not for reading: Results from a longitudinal study. *Reading Research Quarterly, 30*(4), 894-906.

Slavin, R. E. (1995). *Cooperative learning: Theory, research, and practice* (2nd ed.). Boston: Allyn & Bacon, Center for Research on the Education of Students Placed at Risk.

Slavin, R.E., & Madden, N.A. (1995, April). *Effects of Success for All on the achievement of English language learners*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Slavin, R. E., & Madden, N.A. (1999a). Effects of bilingual and English as a second language adaptations of Success for All on the reading achievement of students acquiring English. *Journal of Education for Students Placed at Risk, 4*(4), 393-416.

Slavin, R. E., & Madden, N. A. (1999b). *Success for All/Roots & Wings: 1999 summary of research on achievement outcomes*. Baltimore: Johns Hopkins University.

Snow, C. E., Burns, M. S., & Griffin, P. (1998). *Preventing reading difficulties in young children.* Washington, DC: National Academy Press.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21,* 360-407.

Stone, C. A. (1998). Moving validated instructional practices into the classroom: Learning from examples about the rough road to success. *Learning Disabilities Research & Practice, 13*(3), 121-125.

Stringfield, S., & Teddlie, C. (1991). School, classroom, and student level indicators of rural school effectiveness. *Journal of Research in Rural Education, 7*(3), 15-28.

Tinajero, J. V., & Schifini, A. (1997). *Into English.* Carmel, CA: Hampton-Brown Books.

Thomas, W. P., & Collier, V. P. (1997). *School effectiveness for language minority students.* Washington, DC: National Clearinghouse for Bilingual Education.

Thomas, W.P., & Collier, V.P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement.* Santa Cruz, CA: Center for Research on Education, Diversity & Excellence.

Turner, J. C., & Meyer, D. K. (2000). Studying and understanding the instructional contexts of classrooms: Using our past to forge our future. *Educational Psychologist, 35*(2), 69-85.

Turner, J. C., Meyer, D. K., Cox, K. E., Logan, C., DiCintio, M., & Thomas, C. T. (1998). Creating contexts for involvement in mathematics. *Journal of Educational Psychology, 90*(4), 730-745.

U.S. Census Bureau. (2000). *(NP-T5) Projections of the total resident population by five-year age groups, race, and Hispanic origin, and nativity: Middle series, 1999-2100.* Washington, DC: Author.

Walberg, H. J. (1992). The knowledge base for educational productivity. *International Journal of Educational Reform, 1*(1), 5-15.

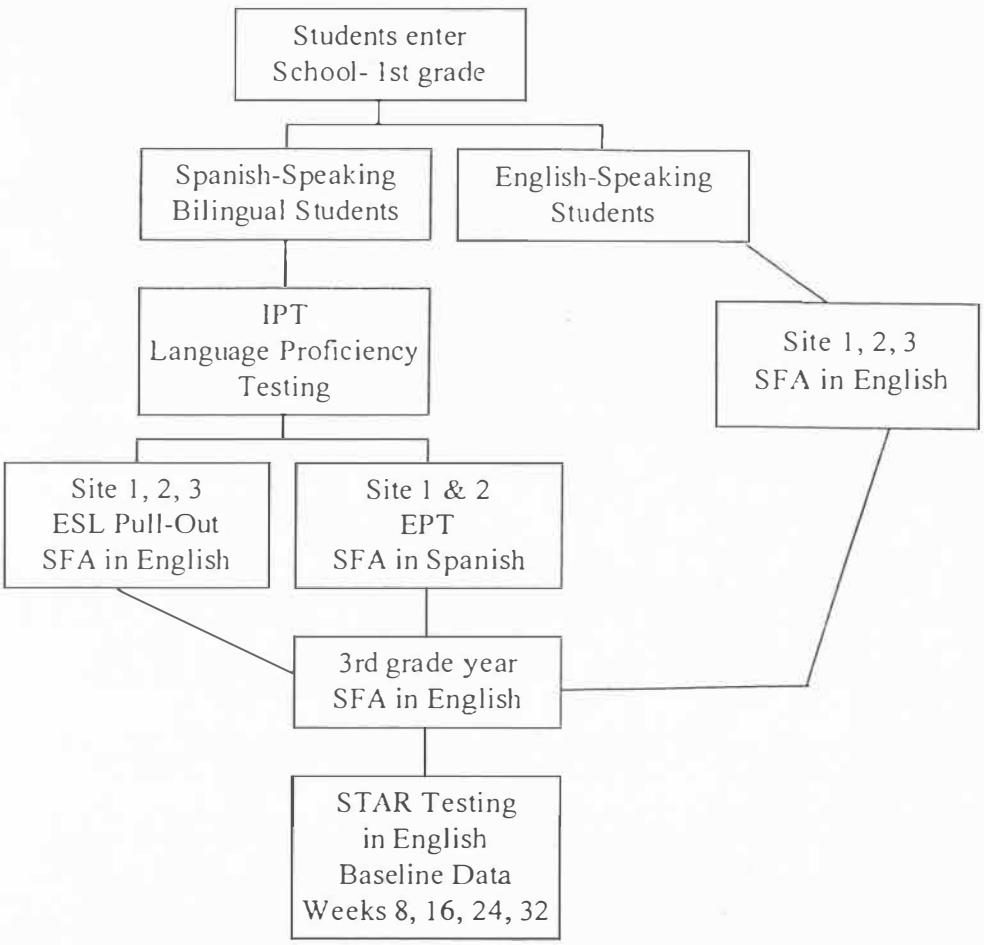
Willig, A. C. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*(3), 269-317.

Woodcock, R. W. (1984). *Woodcock Language Proficiency Battery.* Allen, TX: DLM.

APPENDICES

Appendix A:

Flow Chart Representing Research Design



Appendix B:

English Oral IPT Protocol

Sample Items

- What is this? This is a_____? (shown picture of an apple)
- Listen carefully, then say exactly what I say.

Please take us to the zoo.

I study hard when I go to school.

- Tell me all the days of the week.
- What did you do during lunchtime after you eat? Tell me in a sentence.
- Now, think about what you're going to do tomorrow. Tell me two things you will do tomorrow.

Appendix C:

High-Stakes Tests Compared with STAR Test at Third Grade

California Achievement Test	0.85
Degrees of Reading Power	0.71
Gates-MacGinitie Reading Test (Spring)	0.62
Indiana Statewide Testing for Educational Progress	0.81
Iowa Test of Basic Skills	0.81
Metropolitan Achievement Test	0.73
Missouri Mastery Achievement Test	0.44
MRT	0.81
Stanford Achievement Test (Spring)	0.79
Terra/Nova	0.78

Appendix D:

Data Collection Sheet

School Year 20??-?? <School>

ID# # of years at school Birthday Retained Y/N LEP Y/N if Y, initial IPT if Y, ESL or Spanish SFA

Appendix E:

Sample Growth Trajectory

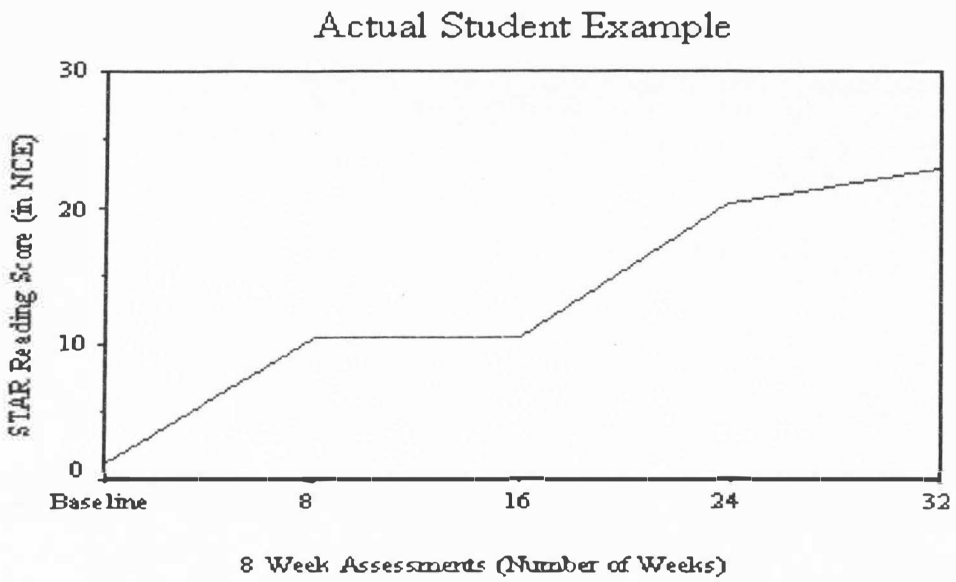


Figure 2. Sample Growth Trajectory