DISSERTATION

EFFECTS OF AN ELEMENTARY TWO WAY BILINGUAL SPANISH-ENGLISH IMMERSION SCHOOL PROGRAM ON JUNIOR HIGH AND HIGH SCHOOL STUDENT ACHIEVEMENT

Submitted by

Luis Diego Vega

School of Education

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2014

Doctoral Committee:

Advisor: R. Brian Cobb

George A. Morgan
Dan Robinson
Tony S. Zimmerman

ABSTRACT

EFFECTS OF AN ELEMENTARY TWO WAY BILINGUAL
SPANISH-ENGLISH IMMERSION SCHOOL PROGRAM
ON JUNIOR HIGH AND HIGH SCHOOL STUDENT ACHIEVEMENT

This study explores the effects of a Two-Way Bilingual Immersion (TWBI) program on language majority and minority students. The fundamental hypothesis was that the process of receiving instruction in two languages (English and Spanish) throughout elementary school (i.e., attendance at a TWBI school) would help the native Spanish-speaking students and not have a negative effect on the native English-speaking students in the performance of core academic areas (reading, mathematics, writing), and that this beneficial effect would carry through Junior High and High School in which instruction was delivered through a "business as usual" English-only model.

This is a longitudinal quasi-experimental study with an ex post facto, non-randomized, matched-pairs design. A multi-level matching procedure was used to match students from the TWBI elementary school (treatment group) with comparable students from throughout the school district (control group) beginning in third grade. Eleven annual cohorts of students from the treatment school were matched on a student-by-student basis on seven variables – cohort year, student's primary language, years of enrollment in the program, ethnicity, gender, socioeconomic status, and 3rd grade performance – with comparable students from within the school district. These eleven cohorts of 3rd graders were then tracked to the end of elementary school, middle and high school and measured on their reading, writing, and math achievement scores at each year. ACT scores were also collected in 11th grade.

We found that students who graduated from the TWBI program had significantly higher CSAP reading, writing and math scores at the end of their elementary school when compared with their matched pairs. We also observed a consistent main effect on program type over time across all three outcome domains, indicating the strength and breadth of the intervention across Junior High and High School.

Native Spanish-speaking students who graduated from the TWBI program achieved significantly better in reading and math, and somewhat better in writing across Junior High through 10th grade than the matched control group. Native English-speaking students who graduated from the treatment program achieved as well as their matched counter parts in writing and math across Junior High through 10th. Furthermore, in the reading area, native English-speaking students who graduated from the treatment program achieved significantly better than their matched controls.

We found that the overall program main effect was small in all three CSAP areas (reading, writing, and math), with at least three interesting trends. First, effect sizes (ESs) tended to be higher for native Spanish-speaking than for native English-speaking students in all three domains, and especially in grades 8, 9 and 10. Second, ESs tended to get bigger for native Spanish-speaking students and smaller for native English-speaking students across Junior High and High School (time) in all three domains. Third, ESs for native Spanish-speaking students in math were the biggest ones at each grade level, with only the exception of 9th grade. Also, math ESs for Spanish-speaking students were bigger than reading and writing ESs for this language group. ESs for native Spanish-speaking students in math were bigger than all ESs for English-speaking students. The treatment program had its biggest effect in the math area for native Spanish-speaking students.

Results also indicate that all students who attended the TWBI program performed better in ACT English, reading, and math scores when compared with their matched pairs. ACT Reading scores were significantly higher for native Spanish-speaking students than for their matched pairs (d = .72), but this was not the case for English, math and science. Native English-speaking students from the treatment group performed equal to or better than their matched counterparts. Furthermore, students from the treatment program obtained mean ACT scores significantly higher than the control group in English (d = .28), reading (d = .36), and math (d = .35) but not science (d = .22). Effect sizes were medium and large for native Spanish-speaking students in English and Reading while they were small to medium for native English-speaking students in these areas, a pattern that is similar to the one that was observed in grades 6 to 10.

Findings suggest consistent support for the two-way immersion program over matched control students across all three achievement areas in Junior High and in three of the four areas evaluated in High School. It appears the greatest effect for native English speakers may be in reading, while native Spanish speakers may benefit more in writing and mathematics. Limitations to generalizability and causal inferences due to the small sample sizes and inherent weaknesses of the research design are noted.

The analysis of attrition revealed that native Spanish speakers from the TWBI program were more likely to stay in the school district than native Spanish speakers from other programs. This was an unexpected but important finding. It could be possible that native Spanish speakers who attended the TWBI program received the benefits of a coherent and theory-based program that successfully helped them improve their academic achievement and allowed them to pursue and navigate their secondary level of instruction.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DEFINITION OF TERMS

The definitions of terms to be used in this report are as follows.

- Additive bilingualism- bilingual development in which there is substantial support for continued L1 development as the individual acquires L2.
- ACT- standardized test for high school achievement and college admissions in the U.S. produced by ACT, Inc. The ACT has historically consisted of four tests: English, Mathematics, Reading, and Science Reasoning.
- CSAP- Colorado Student Assessment Program
- TWBI- Two-way Bilingual Immersion Program
- ESL- English as a Second Language
- ELL- English language learner – The term "English Language Learner," or ELL, refers to any individual who is learning English and for which English is not the native tongue. In state and federal regulations, they are generally referred to as limited English proficient (LEP) students.
- NES- native English-speaking student
- NSS- native Spanish-speaking student
- L1- an individual's first language
- L2- an individual's second language
- Subtractive bilingualism-bilingual development in which use of the majority culture's language (L2) is required or strongly encouraged and thought to replace L1.

CHAPTER 1: INTRODUCTION

The ethnic and racial composition of the United States of America (U.S.) population is rapidly changing. High levels of immigration along with high fertility rates in minority groups have contributed significantly to such change. In 2010, 50.5 million or 16.3 percent of the inhabitants of the U.S. were of Hispanic or Latino origin (Enis, Ríos-Vargas, & Albert, 2011). Along with this change, the student population is also experiencing a transformation. There were 5.3 million students classified as English language learners (ELLs) enrolled in the K-12 public schools in the 2008-2009 academic year. They represented 10.8 percent of the student body. ELL students are the fastest-growing segment of the student population. About 79 percent of ELL students in the U.S speak Spanish as their native language (U.S. Department of Education, 2011).

In an effort to respond to and better serve the needs of the ELL students, several models of bilingual education have been proposed: Submersion, English as a Second Language (ESL), Shelter English Instruction, Newcomer Programs, Transitional Bilingual Education (TBE), Developmental Bilingual Education (DBE), Foreign Language Immersion (FLI), and Two-Way Bilingual Immersion (TWBI).

- Submersion. This model consists in placing ELL students in regular English-speaking classrooms with minimal instruction in the actual mechanics of English.
- English as a Second Language. In this model, students are "pulled out" of some other classes in order to receive an English-as-a-second-language class.
- Shelter English instruction. In this model, ESL and content area classes are combined, and taught by an ESL-trained subject area teacher.

- Newcomer Programs. These programs support rapid English acquisition, and are often located off-campus. They are designed for non-English-speaking students in middle and high school who recently arrived to the U.S.
- Transitional Bilingual Education. It is the most common form of bilingual education for ELL students in the U.S. (Genesse, 1999). TBE programs seek to achieve basic oral English proficiency within 2 years and to mainstream students to an all-English program within 3 years.
- Developmental Bilingual Education. It is an enrichment program that educates ELL students using both English and their first language for academic instruction.
- Foreign Language Immersion. In FLI programs, teachers use a second/foreign language as the medium of academic instruction and social interaction with native-English-speaking majority group students. The second or foreign language is used for at least 50% and up to 100% of academic instruction during the elementary or secondary grades.
- Two-Way Bilingual Immersion. TWBI programs provide integrated language and academic instruction for native English-speaking students and native-speaking students of another language with the goals of high academic achievement, first and second language proficiency, and cross-cultural understanding.

More detail about these programs and their current research base will be provided in Chapter 2.

The intent of this study was to investigate the effects of a TWBI program on reading, writing, and math achievement on native Spanish-speaking students (NSS) and native English-speaking students (NES). Using a longitudinal sound comparative methodology, this study contributes to base of knowledge on bilingual education in the U.S.

This study is part of longitudinal research conducted on a TWBI program in Northern Colorado. The first findings have been already published (Cobb, Vega & Kronauge, 2006). New data collection on Colorado Student Assessment Program (CSAP) scores and Achievement College Test (ACT) scores made it possible to further investigate the effects of this program on students' academic achievement from Kindergarten through 12th grade.

Study Context

Research findings in bilingual education are mostly unambiguous regarding the positive effects of bilingualism on children´s awareness of language and cognitive functioning (Bialysotck, 2001; Cummins, 2000); nevertheless, bilingual education remains as a highly debated issue in the U.S. and in other parts of the world. Controversy about the appropriateness and effectiveness of bilingual education has been a major focus of public debate (Bekerman, 2005).

The debate around bilingual education is not only about instructional methods, but about much larger philosophical arguments over language rights, cultural inclusion, and political representation. This type of education has been the focus of controversy because it raises questions about national identity, federalism, power, ethnicity, and pedagogy.

In the U.S., bilingual education must be linked to the historical context of immigration as well as political movements such as civil rights, equality of educational opportunity, affirmative action and melting pot policies (Baker, 2006). Traditionally, those in favor of bilingual education are language specialists, Mexican American activists, civil rights advocates, language minorities, intellectuals, teachers, and students. They are ideologically opposed to the assimilationist philosophy in the schools, to the structural exclusion and institutional discrimination of minority groups, and to limited school reform. On the other hand, the opponents of bilingual education,

have been, at different points in time, conservative journalists, politicians, federal bureaucrats, Anglo parent groups, school officials, administrators, and special-interest groups (such as U.S. English). They favor assimilationism of ethnic minorities, and limited school reform (San Miguel, 2004).

Changing Demographics in the U.S.

The composition of the U.S. population has experienced significance changes in the recent past. For example, while the White population grew in every decade throughout the 20th century, its share of the total U.S. population did not follow this same pattern. At the beginning of the century, just 1 out of 8 people who inhabit U.S. was of a race other than White. At the end of the century, the proportion was 1 out of 4. The latest census data reveals that in 2010, 37.6 percent of the inhabitants of the U.S. indicated that their race was other than White (Humes, Jones, & Ramirez, 2011). The population of the U.S. is diverse and will become even more diverse in the next decades.

A review of information available between 1980 and 2010 indicate that over the last three decades the Hispanic population has more than tripled in size in the U.S. In 1980, there were 14.6 million Hispanics in the U.S. In 2000, there were 35.3 million. This figure went up to 50.5 million or 16.3 percent of the total population in 2010. High levels of immigration along with high fertility rates have contributed significantly to such rapid growth. (Humes, Jones, & Ramirez, 2011; Hobbs & Stoops, 2002). Furthermore, the Hispanic population increased by 15.2 million between 2000 and 2010 (see Figure 1), accounting for over half of the 27.3 million increase in the total population of the U.S. Between 2000 and 2010, the Hispanic population grew by 43 percent, which was four times the growth in the total population at 10 percent. Within the Hispanic group, the Mexican origin population increased by 54 percent and had the

largest numeric change (11.2 million), growing from 20.6 million in 2000 to 31.8 million in 2010. Accurate data on the Hispanic group in the U.S. is available only since 1980 due to the format of the questions regarding race and ethnicity that were used in census prior to that year (Humes, Jones, & Ramirez, 2011).

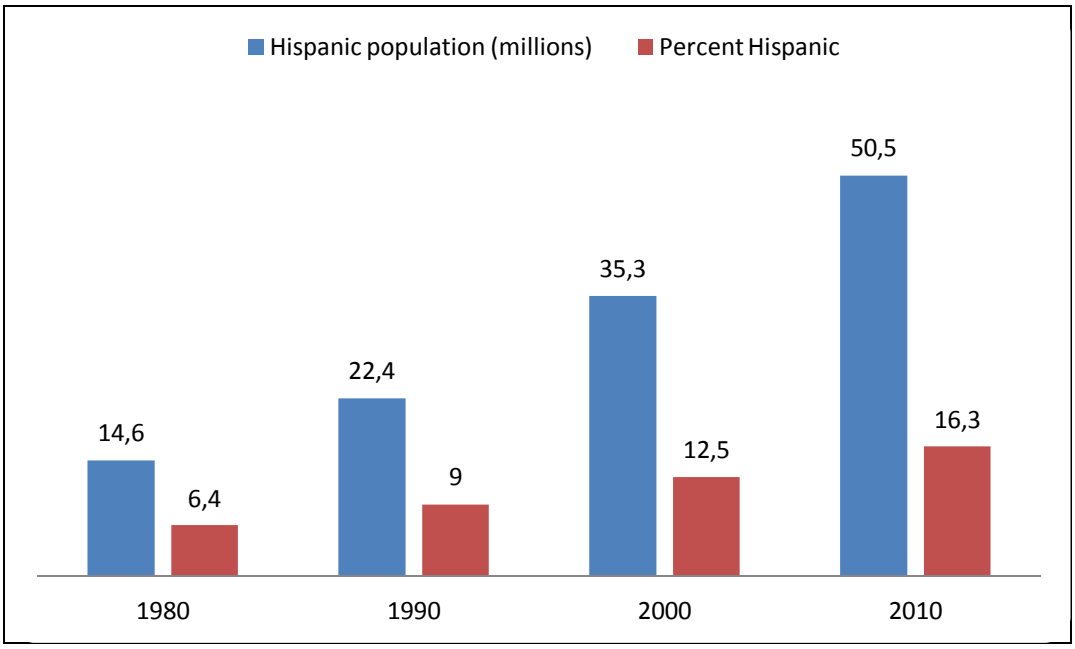

Figure 1. Hispanic population and percent Hispanic of total population: 1980 to 2010

In a similar way, the makeup of students in PreK–12 classrooms across the U.S. has become increasingly more diverse. During the 2008–09 school year, ELLs represented 10.8 percent of the K–12 public school enrollment, or more than 5.3 million students. In fact, ELLs are the fastest-growing segment of the student population, with their growth highest in grades seven through twelve (NCELA, 2011). About 79 percent of ELLs in the U.S. are NSS; a much lower percent has Chinese, Vietnamese, Hmong, or Korean as their native language. As a group, ELLs are more likely to come from lower economic and educational backgrounds. The majority of ELLs enrolled in the public schools in U.S. in 2008-09 were born in the U.S (65.3 percent), followed by those who were born in México (18.6 percent). The rest of ELLs (16,1%) came

from many different countries like China (1.2 percent), El Salvador (1 percent), Korea (1 percent), Philippines (.09 percent), Dominican Republic (.08 percent), and Vietnam (.08 percent) (Soltero, 2011).

Statement of the Research Problem

Schools in the U.S. are responsible for educating the nearly 6 million school-aged children in the U.S. whose proficiency in English varies widely. The No Child Left Behind Act (2001) requires schools to help children attain English proficiency and high levels of academic attainment as quickly and effectively as possible (U.S. Department of Education, 2002). In addition, native English speakers are becoming more interested in acquiring a second language in school as is evidenced by the growing number of two-way bilingual education programs in the U.S. Prior to 1995, there were less than 90 programs in the U.S. Recently there are over 400 two-way bilingual education programs in the U.S. (Center for Applied Linguistics, 2011).

The effectiveness of bilingual education continues to be highly debated in the U.S., despite a growing body of research that demonstrates its validity. Results of at least five meta-analyes point in the direction of the advantage of bilingual education over all-English programs (Greene, 1997; MacField, 2002; Rolstad, Mahoney, & Glass, 2005; Slavin & Chegun, 2005; Willig, 1985). The question should be, not if bilingual education is effective, but "What types of bilingual program work"? Here, more evidence is needed. For example, after reviewing selected bilingual education studies, McField (2002) concluded that programs designed along principles hypothesized to underlie ideal bilingual programs were more effective. But very few such comparisons were possible. Only one "strong" program and four "weak" programs could be analyzed in this way. In particular, the effects of TWBI programs on language majority and minority students need further explanation because studies of these types of programs are

lacking. In addition, more methodologically sound studies are needed that examine achievement trends over six or more years and control for important student background characteristics.

The basis for conducting this study of a TWBI education program was: 1) to investigate the effects of the program on reading, writing, and math achievement in this study population, 2) to contribute to the base of knowledge on bilingual education in the U.S. and 3) to promote the use of methodologically sound comparative research in the bilingual education field.

Research Hypotheses

The fundamental hypothesis was that the process of receiving instruction in two languages (English and Spanish) throughout elementary school (attendance at a TWBI school) would help the native Spanish speaking students in the school and not have a negative effect on the native English speaking students in the school in core academic areas (reading, mathematics, writing), and that this beneficial effect would carry through junior high and high school in which instruction was delivered through a "business as usual" English-only model. This fundamental hypothesis shaped the following research hypotheses:

1. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in reading, mathematics, and writing at the end of 6^{th}, 7^{th}, 8^{th}, 9^{th}, and 10^{th} grades than will a matched control group as measured by the Colorado Student Assessment Program test battery.
2. Those native Spanish speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in each of these achievement domains and each of these five years than a matched control group.

3. Those native English speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in each of these achievement domains and each of these five years as a matched control group.

4. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group as measured by the American College Testing test battery.
5. Those native Spanish speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group.
6. Those native English speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in English, reading, mathematics, and science in 11th grade as a matched control group.

Specific mean comparisons among groups reflecting these research hypotheses are presented in Table 1.

Study Significance

The rationale for this research stems from the opponents assertion that for students whose primary language is English (NES), learning elementary school content in math and language arts *in both Spanish and English* can put those students at a disadvantage compared to students educated in only-English programs when they get to junior high and high school due to the fact that they have to learn a second language at the same time that they have to learn math and reading content. The rationale for this research also derives from the ongoing political debate

Table 1.

Mean Comparisons Reflecting Research Hypotheses

Grade Level	Treat-ment Group	Speak-ing Group	CSAP Test Areas						ACT Test Areas							
			Reading		Math		Writing		Reading		Math		English		Science	
			H_1	$H_{2\&3}$	H_1	$H_{2\&3}$	H_1	$H_{2\&3}$	H_4	$H_{5\&6}$	H_4	$H_{5\&6}$	H_4	$H_{5\&6}$	H_4	$H_{5\&6}$
6th-10th Grade	TWBI	NES	M>	M=	M>	M=	M>	M=								
		NSS		M>		M>		M>								
	Cont.	NES	M<	M=	M<	M=	M<	M=								
		NSS		M<		M<		M<								
11th Grade	TWBI	NES							M>	M=	M>	M=	M>	M=	M>	M=
		NSS								M>		M>		M>		M>
	Cont.	NES							M<	M=	M<	M=	M<	M=	M<	M=
		NSS								M<		M<		M<		M<

Note: M stands for "Mean"

H_1 and H_4 stand for the main effect of program in hypothesis 1 and 4, respectively

$H_{2\&3}$ and $H_{5\&6}$ stand for the main effect of program in hypothesis 2 & 3 and 5 & 6, respectively

TWBI stands for "Two-way Bilingual Immersion"

NES stands for "Native English Speaker"

NSS stands for "Native Spanish Speaker"

Cont. stands for "Control Group"

about what is the best way to educate NSS students. This research informs this debate from a longitudinal perspective by tracking several cohorts of TWBI students (NES and NSS), matching them with students who attended regular elementary schools programs, and comparing their academic achievement at a different points in time throughout junior high and high school.

Delimitations and Limitations

There are a number of methodological challenges associated with research on TWBI programs that make it difficult to pinpoint definitive findings. The first one is lack of randomization. This longitudinal study was done on a TWBI program that is voluntary; therefore, self-selection may influence student outcomes. In other words, if students of the TWBI program are found to do better than their peers in other programs, it is difficult to know if this is because of the effects of the program itself, or due, at least in part, to inherent differences among the student population and their families who decided to opt into the TWBI program. However, the relatively sophisticated matching process used in this research, can take some of the uncertainty out of this, but not at the level that would have been the case had random assignment to groups been used.

Another limitation is related with differences between groups in socioeconomic status. There was a significant difference between NSS students and NES students enrolled in the program. NSS students were more likely to come from homes where there is poverty and where parents have limited formal schooling; however NES students were more likely to come from homes that are solidly middle class and where parents have substantial formal education. This difference in the backgrounds of the two groups of students makes internal comparisons of English versus Spanish student performance difficult, as the students frequently differ by more than just native language.

Finally, geographical mobility is a variable that can influence the results. Only students who stayed in the school district where the TWBI program is located for the duration of the study were included in the sample. Only students who completed at least four years in the TWBI program were included in the sample. Thus, students that left the district or did not complete the entire program were excluded from the study. The effects of this geographical mobility are not clear and may produce biased results. The study was confined to one TWBI program in Northern Colorado. This program may not represent all TWBI programs been implemented in the country. The researcher was the school counselor at the elementary school where the study was conducted during two of the initial years included in the study.

CHAPTER 2: REVIEW OF THE LITERATURE

The intent of this chapter is to present a review of the most relevant literature related with two-way bilingual immersion (TWBI) programs and their effectiveness. It includes a brief presentation of the history of bilingual education in the U.S. explaining each of its identified periods: Permissive, Restrictive, Opportunist, and Dismissive. This section is followed by a characterization of the eight most common models of bilingual education utilized in the U.S.: Submersion, English as a Second Language, Shelter English Instruction, Newcomer Programs, Transitional Bilingual Education, Developmental Bilingual Education, Foreign Language Immersion, and TWBI. Finally, types and effectiveness of the TWBI models are presented.

History of Bilingual Education in the U.S.

The historical origins of bilingual education in the U.S. can be found well before 1963 when the first two-way bilingual program was implemented in Coral Way Elementary School in Dade County, Florida. Bilingual education in the U.S. has moved through constant change in ideology, preference and practice. This history should be linked to the historical context of immigration and political movements like civil rights or equality of education opportunity (Baker, 2006).

A review of the history of bilingual education shows that language ideology in the U.S. has shifted according to changing historical events. It has not maintained a stable course (Ovando, 2003). As Paulston (1992, p. 80) observes: “unless we try in some way to account for the socio-historical, cultural, and economic-political factors which lead to certain forms of bilingual education, we will never understand the consequences of that education”

Four distinctive periods have been identified in the history of bilingual education in the U.S.: Permissive, Restrictive, Opportunist, and Dismissive.

Permissive period: 1700s–1880s. A climate of linguistic tolerance, especially for those from Northern Europe, was present in the U.S. during the 18th and 19th centuries. Linguistic diversity was often accepted and the presence of different languages was frequently encouraged through religious services, community newspapers, and in both private and public schools (Baker, 2006).

During the second half of the 19th century, bilingual or non-English-language instruction was provided in some form in many public and private schools: German in Pennsylvania, Maryland, Ohio, Indiana, Illinois, Missouri, Nebraska, Colorado, and Oregon; Swedish, Norwegian, and Danish in Wisconsin, Illinois, Minnesota, Iowa, North and South Dakota, Nebraska, and Washington; Dutch in Michigan; Polish and Italian in Wisconsin; Czech in Texas; French in Louisiana; and Spanish in the Southwest. A number of states passed laws that authorized bilingual education in this period (Leung, 2008). The rationale was that immigrant communities could still maintain their ancestral roots while actively participating in civil life.

Although this period can be characterized as permissive, it is important to keep in mind that 19th-century education did not actively promote bilingualism. Rather, a policy of linguistic assimilation without coercion seemed to prevail (Ovando, 2003).

Restrictive period: 1880s–1960s. The 1880s were a turning point in the historical development of linguistic and immigration restrictionism as a number of repressive policies appeared. Beginning in this decade, the government attempted to repress Indians by issuing restrictive policies that contained them in their reservations. The American Protective Association was one of the several organizations that promoted English-only instruction. By the 1880s, the Bureau of Indian Affairs implemented a policy of forced Anglicisation for Native Americans sending Indian children to boarding schools. Such policies did not succeed in

eradicating the children's native languages, but it did instil in them a sense of shame that guaranteed the exclusive use of English for future generations (Nieto, 2009).

Increasing fear about the importation of foreign ideologies into the U.S. resulted in a call for all immigrants to be assimilated into one cultural and linguistic mold. For example, the Naturalization Act of 1906 required all immigrants to speak English in order to be eligible to start their process of naturalization (Ovando, 2003). This justification of the imposition of English was based on the explicit connection between English and U.S. national identity.

The declaration of war on Germany in World War I served as a catalyst for English over German instruction. At this time, the previous tolerance toward German speakers turned to hostility. Therefore, most school districts eliminated German instruction from their curriculum because it was seen as anti-American. (Leung, 2007)

The predominant approach to educating language-minority students in the U.S. during this period was the sink-or-swim method, also known as submersion. Most educators and policy makers felt that it was up to the language-minority students, not the schools, to make the linguistic, cultural, and cognitive adjustments necessary to achieve assimilation into American society. This push for homogeneity became a well-established pattern within schools during the first half of the 20th century (Ovando, 2003). By 1923, the legislatures of 34 states had dictated English-only instruction in all private and public primary schools.

Although the foreign language instruction was in the direction of monolingualism during the period, the Supreme Court declared that Nebraska's prohibition against the teaching of foreign languages in elementary schools (Meyer v. Nebraska, 1923) was unconstitutional based on the 14th Amendment (Baker 2006).

Opportunist period: 1960s–1980s. World War II and the cold word served as a wake-up call for addressing the inadequacies in foreign-language instruction. In 1957, the former Soviet Union launched their *Sputnik* into the space. In the U.S., it led to general discussions and questioning on its ability to compete in an increasingly international world and on the quality of the education in the country. Because language, math, and science skills were essential for military, commercial, and diplomatic endeavors, these subjects became a high priority in the national defense agenda during the cold war period. A new consciousness aroused about the need for foreign language instruction. This led to the creation of the National Defense Education Act in 1958. One of the act's primary goals was to raise the level of foreign-language education in the U.S.. Although this Act promoted much-needed improvement in the teaching of foreign languages, it did not alter the linguistically disjointed tradition of the U.S. On one hand, the country was encouraging the study of foreign languages for English monolinguals, at great cost and with great inefficiency. At the same time, it was destroying through monolingual English instruction the linguistic gifts that children from non-English-language backgrounds bring to our schools (Ovando, 2003).

The 1906 Naturalization Act was revoked by the 1965 Immigration and Nationality Act, which eliminated racial criteria for admission, expanding immigration especially from Asia and Latin America. The Act also emphasized the goal of 'family unification' over occupational skills. This encouraged increased immigration by Mexicans in particular (Baker, 2006). The lack of access to a meaningful education hindered the possibility of full participation in society for these non-English-speaking students and blocked their upward mobility. Both facts motivated Congress to pass the Bilingual Education Act of 1968, also known as Title VII of the Elementary and Secondary Education Act. This Act provided funding to establish bilingual programs for

students who did not speak English and who were economically poor. The Bilingual Education Act has been considered the most important law in recognizing linguistic minority rights in the history of the U.S. (Nieto, 2009).

The first two-way bilingual education program in the U.S. was established at Coral Way Elementary School in Dade County, Florida in 1963. The program became an option for children of exiled Cubans arriving to Florida after Fidel Castro's Cuban Revolution in 1959. Coral Way's success with bilingual education stimulated other bilingual immersion programs in Florida and other parts of the country (Ovando, 2003).

The next important event in the rebirth of bilingual education was the 1974 Supreme Court case *Lau v. Nichols* (414 U.S. 5637). The *Lau* decision was the result of a class action suit representing 1,800 Chinese students who alleged discrimination on the grounds that they could not achieve academically because they did not understand the instruction of their English-speaking teachers. This ruling reinforced the mandate that it was the school district's responsibility to provide the necessary programs and accommodations to children who did not speak English. Basing their unanimous decision on the 1964 Civil Rights Act, the justices concluded that the responsibility to overcome language barriers that impede full integration of students falls on the school boards and not on the parents or children; otherwise, there is no real access for these students to a meaningful education (Nieto, 2009). The *Lau* decision has had an enormous impact on the development of bilingual education in the U.S.. The *Lau* verdict abolished the sink-or-swim practices of the past and led to the passage of the Equal Educational Opportunities Act in August 1974. With this act, Congress required every school district to take appropriate action to overcome language barriers that impede equal participation by its students in its instructional programs.

Many opportunities for the development of bilingual education were crystallized during this period, thus affirming the civil rights of language-minority students. However, despite its growth, bilingual education continued to remain controversial during this period, as evidenced by the 1972 finding of the U.S. Commission on Civil Rights that only a very small percentage of language minority students were receiving appropriate bilingual or ESL instruction in California, Arizona, New Mexico, Colorado, and Texas (Ovando, 2003).

Dismissive Period: 1980s–Present. Despite bilingual education was spreading throughout the United Stated, the sentiment against bilingual education regained strength in this period. For example, in 1983, senator S.I. Hayakawa founded U.S. English (http://www.us-english.org/), a non-profit organization that promotes English as the official language of the U.S. and discredits bilingual education. In the eighties, the Reagan administration led a major campaign against bilingual education and in favour of a "back to basics" education. The Reagan administration defined the U.S. as a "nation at risk of balkanization" and blamed non-English speaking communities for such a risk (Crawford, 1989). This trend continued into the 1990s. Political activists across the nation began to press for a return to the sink-or-swim days and the melting pot ideology. Antibilingual education pressure groups such as U.S. English, English Only, and English First began to appear on the scene (Baker, 2006).

In 1994, California voters approved Proposition 187, a ballot initiative designed to sharply curb illegal immigration through strong restrictions on the social and educational services that undocumented immigrants could receive. In 1998, Proposition 227, promoted by multimillionaire Ron Unz, was adopted in California. Proposition 227 sought to impose severe restrictions on native-language instruction for English learners in California. Most bilingual programs were dismantled and substituted with English-only instruction models with English

learners receiving less help than before in their native languages. Similar measures were passed in Arizona in 2000 and in Massachusetts in 2002 (Leung, 2008).

This wave of anti-bilingualism policies reached its peak with George W. Bush's No Child Left Behind Act (NCLB) in 2002. The law did not officially ban bilingual programs, but it imposed a high-stakes testing system that promoted the adoption and implementation of English-only instruction. Furthermore, all references to bilingual education were eliminated in the new legislation (Nieto, 2009). English Only, U.S. English, English First, Proposition 187, Proposition 227, and the proposed riders to federal bilingual funding can be seen collectively as instruments of the politics of resentment toward massive immigration from developing countries in the 1980s and 1990s, especially from Asia and Latin America (Ovando, 2003).

Bilingual Educational Models in the U.S.

A useful framework to classify and better understands the vast diversity of educational models and programs designed to help English language learners (ELL) students in the U.S. is to organize them according to their national or societal goals and their intended outcomes.
In general, national goals are of two types: assimilationist and pluralistic. Each of them reflects ideological and philosophical standpoints. Assimilationist goals seek to assimilate minority language speakers into the majority language and culture. As a result of this process the minority language would become less important or even disappear. The image of a "melting pot", associated with this goal, implies that failure to assimilate may lead to separatism. Pluralistic goals affirm individual and group language rights. They are also associated with support for group autonomy, which may or may not be viewed as a threat to larger group unity.

Outcomes are typically categorized as: subtractive bilingualism and additive bilingualism. Subtractive bilingualism occurs when students lose their first language in the process of

acquiring a second language. Additive bilingualism is what results from a program in which students maintain their first language and acquire a second one (Roberts, 1995).

Subtractive programs include submersion, English as a second language, transitional bilingual education, and newcomer programs. Additive programs, on the other hand, promote bilingualism by incorporating both the minority language and English into the academic setting. Such programs include developmental bilingual education, foreign language immersion education, and two-way immersion. A description of each program model as described by Genesse (1999) and Roberts (1995) is provided below (see also table 1). Some disagreement exists over the classification and labeling of the programs among researchers, politicians, defendants and detractors of bilingual education. Additional labels are provided for each model to minimize misunderstanding of the terms.

Submersion. Also known as Structured English Immersion. This is a program with an assimilationist goal and a subtractive approach that requires minority language students to progress through the same academic content as their English-speaking peers. It consists in placing ELL students in regular English-speaking classrooms that feature little or no instructional modifications and minimal instruction in the actual mechanics of English. Since L1 is not supported, it is frequently lost. This is by far the most widely used format for instruction of ELL students in the U.S.' public schools (Soltero, 2004). In this model, students of the minority language are placed in English-only classrooms and receive little, if any, first-language support and minimal pull-out assistance with English acquisition. According to Roberts (1995), submersion is not a legal option for schools with ELL students; however, oversight and enforcement are lax. Parents of these children have the right to demand the services their children need, but for cultural of other reasons, rarely do it.

English as a second language (ESL). Also known as ESL Pullout. In this model, students are "pulled out" of some other classes in order to receive an English as a second language class. This program is also assimilationist in its goals, and subtractive bilingualism is its usual outcome. It is commonly found in areas with students of a variety of language backgrounds, and in areas where resources are limited. ESL instruction may exist as part of a bilingual education model or as a stand-alone program. As a separate program, an ESL approach is especially practical when students wishing to acquire English represent several different languages. In its most common form, ESL services are provided on a pull-out basis. Students are taught English as a subject matter. Instruction is focused on vocabulary, grammar, and verb usage through drill and practice sessions. When implementing this program, it makes sense to release ELL students from English Language Arts classes to attend their ESL classes. It is less appropriate to take children from content classes or from classes in which they can form friendships with native English-speaking students (NES), such as P.E., music, or art.

Shelter English instruction. Also known as ESL with content-based instruction. In this model ESL and content area classes are combined, and taught either by an ESL-trained subject area teacher or a team. These classes are designed to deliver content area instruction in a form more accessible than the mainstream. They may use additional materials, bilingual aides, adapted texts, and so on to help ELL students acquire the content as well as the language. Sheltered programs are also assimilationist. Students develop their English skills through study of the academic curriculum. This allows them to learn major concepts that are being taught in the classroom as they acquire English comprehension and fluency. Sheltered Instruction can be a program option in itself or an approach used in conjunction with other programs. For example, it can be the method used to teach the English component of transitional bilingual, developmental

bilingual, or two-way immersion programs. Sheltered instructional strategies can also be used to teach content through a second/foreign language to native-English-speaking students in foreign/second language immersion programs.

Newcomer Programs. This model focuses on the unique needs of non-English-speaking middle and high school students who recently arrived with little formal education in their native country. Such students require more assistance than ESL services can provide. Newcomer programs support rapid English acquisition as well as assimilation to the new culture. Programs are often located off-campus and, after 6 months to 2 years, students transfer into the school's bilingual or ESL program. The goals of newcomer programs are to help students acquire beginning English language skills along with core academic skills and knowledge, and to acculturate to the U.S. school system. This program is assimilationist in its goals, and subtractive bilingualism is its usual outcome.

Transitional bilingual education (TBE). Also known as Early Exit Bilingual Education. TBE is the most common form of bilingual education for ELL students in the U.S. (Genesse, 1999). The term “transitional” speaks to the process of students moving gradually from instruction primarily in their first language to instruction entirely in English. Most TBE programs start in kindergarten or first grade. They seek to achieve basic oral English proficiency within 2 years and to mainstream students to an all-English program within 3 years. Typically, students who start the program in kindergarten are placed in an all-English program by the beginning of third grade, and those who start in first grade are placed in an all-English program by the beginning of forth grade. These programs are sometimes referred to as early exit bilingual education, because students exit relatively early in comparison to developmental and TWBI programs, which maintain instruction through the first language throughout the elementary

grades. In the beginning, all content and literacy classes are taught in the native language with a gradual increase in the amount of English used for instruction. The goal of this program is for the student to acquire a level of English proficiency to be mainstreamed into the general education classroom, therefore it is still assimilationist. Because the student is eventually mainstreamed into the English classroom, this model is still considered to be subtractive in nature.

Content instruction through English is often provided in individualized and specially designed programs, sometimes referred to as sheltered instruction. As students acquire proficiency in oral English, the language in which academic subjects are taught gradually shifts from the students' first language to English. The transition typically starts off with math computations, followed by reading and writing, then science, and finally social studies. Once they acquire sufficient English proficiency, TBE students transition to mainstream classes where all academic instruction is in English.

Developmental bilingual education (DBE). Also known as Maintenance Bilingual Education, or Late Exit Bilingual Education. DBE differs significantly from the previous models in both goals and outcomes. DBE is pluralistic in its goals because it promotes bilingualism and biliteracy. It is an enrichment program that educates ELL students using both English and their first language for academic instruction. Languages other than English are seen as resources. Because it promotes the development of two languages, the outcome is additive bilingualism. During the 1960s and 1970s, DBE programs were referred to as maintenance bilingual programs; however this name was dropped to avoid negative political associations linked to the notion of first language maintenance. The term developmental bilingual education was first introduced in Title VII of the 1984 Elementary and Secondary Education Act to emphasize the importance of supporting the long-term linguistic, academic, and cognitive development of ELL students

(Genesse, 1999). DBE is a kind of one-way program that includes only or primarily language minority students.

Most DBE programs initially begin with kindergarten or first grade and add one grade each year. They teach regular academic subjects through both English and the students' native language for as many grades as the school district can support, ideally through the end of secondary school. DBE programs aim to promote high levels of academic achievement in all curricular areas and full academic language proficiency in the students' first and second languages. They emphasize the cognitive and academic richness of exploring knowledge across academic domains from multiple cultural perspectives and using two languages. DBE programs provide English language learners with academic instruction in their first language as they learn English. Sheltered instructional techniques are the preferred method of delivering academic instruction. Development of the students' first language is seen as not only feasible but also desirable. It seeks to overcome the perceptions of some school personnel that use of the students' first language is only remedial, serving simply as a bridge to English language development.

Foreign language immersion (FLI). Also known as Second Language Immersion or Heritage Language Immersion. FLI programs are designed for students who come to school speaking the majority language—English in the case of the U.S. They can serve the educational aspirations of English-speaking students who are members of cultural minority groups that wish to promote acquisition of their indigeneous or heritage language–for example, Chinese, German, Navaho, or Hawaiian. They are not intended for ELL students.

In FLI programs in the U.S., teachers use a second/foreign language as the medium of academic instruction and social interaction with native-English-speaking majority group

students. The second or foreign language is used for at least 50% and up to 100% of academic instruction during the elementary or secondary grades.

Immersion is distinctive as a method of foreign/second language instruction because it uses academic content as the medium for second language teaching rather than focusing instruction directly on the teaching of second language skills (Genesee, 1999). Thus, in immersion programs, a great deal of foreign/second language learning occurs incidentally, as students and teachers use the second language to interact with each other about academic content and social matters in school. In this way, learning the second language is similar to the way children learn their first language.

TWBI. Also known as Two Way Immersion Education, or Dual Language Immersion. Despite similar characteristics among the dual language programs, and widespread agreement about the success of these programs, there is not the same agreement about what the programs should be called: *dual language education, two-way bilingual education, two-way immersion, dual immersion*, and *enriched education* are terms used by various scholars (Gómez, Freeman &Freeman, 2005).

TWBI programs provide integrated language and academic instruction for NES students and native-speaking students of another language with the goals of high academic achievement, first and second language proficiency, and cross-cultural understanding. Thus, TWBI programs have a pluralistic goal and additive bilingualism is its usual outcome. In TWBI programs, as in other immersion programs, language learning is integrated with content instruction. Academic subjects are taught to all students through both English and the other language. As students and teachers interact socially and work together to perform academic tasks, the students' language abilities are developed along with their knowledge of academic subject matter. Students in dual

language environments have the added advantage of interacting with peer models of the second language. Both languages and cultures receive equal value and affirmation. This form of additive immersion promotes language development for all students in both the primary and secondary languages. A more detail description of TWBI is presented in the following section.

Main Characteristics of TWBI Programs

Theory on TWBI as a way of teaching. The theoretical foundation for TWBI is based on research findings concerning both first and second language acquisition. Bilingual education research indicates that academic knowledge and skills acquired through one's first language make it easier for acquisition of related knowledge and skills in another language (Collier & Thomas, 2004; Greene, 1997; Willig, 1985). When instruction through the first language is provided to language minority students along with balanced second language support, these students attain higher levels of academic achievement than if they had been taught in the second language only (Collier, 2004; Lanauze & Snow, 1989).

Research indicates that English is best acquired by students with limited or no proficiency in English after their first language is firmly established. Specifically, strong oral and literacy skills developed in the first language provide a solid basis for the acquisition of literacy and other academic language skills in English (Slavin, & Cheung, 2005; Genesee, 1987; Edelsky, 1982; Eisterhold-Carson, Carrell, Silberstein, Kroll, & Kuehn, 1990; Lanauze & Snow, 1989). Moreover, common skills that underlie the acquisition and use of both languages transfer from the first to the second language, thereby facilitating second language acquisition.

Research on immersion programs for language majority students (those who are native speakers of English) has shown evidence that language majority students can maintain grade-level academic achievement and English literacy skills, despite receiving most of their

instruction in a second language. They can also develop advanced levels of second language proficiency without compromising their academic achievement or first language development (Genesee, 1987; Swain & Lapkin, 1991).

Finally, language is learned best by all students when it is the medium of instruction rather than the exclusive focus of instruction. In TWI settings, students learn language while exploring and learning academic content because there is a real need to communicate.

In other words, research indicates that additive bilingual instruction models can be effective for both language minority and language majority students, because they enable the development of language and literacy in both the native language and a second language without diminishing academic achievement (Genesee, 1999).

Defining criteria and goals of TWBI programs. Two-way immersion education is a dynamic form of education that holds great promise for developing high levels of academic achievement, bilingualism and biliteracy, and cross-cultural awareness among participating students. At the same time, because it involves the provision of instruction in two languages to integrated groups of students, it is a complicated and challenging model to implement effectively.

Howard and Christian (2002) proposed three defining criteria of TWBI programs:

- The programs must include fairly equal numbers of two groups of students: language majority students, who in the U.S. are native English speakers; and language minority students, who in the U.S. are native speakers of another language, such as Spanish, Korean, or Chinese. Two-way immersion education is distinct from other forms of dual language education (such as developmental bilingual education or foreign language immersion), because it is two-way in two ways: two languages are used for instruction,

and two groups of students are involved, including native English speakers and language minority students from a single language background, usually Spanish.

- The programs are integrated, meaning that the language majority students and language minority students are grouped together for academic instruction for all or most of the day.
- The programs provide core academic instruction (i.e., content and literacy courses) to both groups of students in both languages. Depending on the program model, literacy instruction may not be provided to both groups in both languages initially, but by about third grade, all students are typically receiving literacy instruction in both languages.

In addition, Howard and Christian (2002), recommend that all TWBI programs must comply with the following goals:

- Students will develop high levels of proficiency in their first language (L1). This means that the language minority students will develop high levels of speaking, listening, reading, and writing ability in their native language (e.g., Spanish) and native English speakers will develop high levels of speaking, listening, reading, and writing ability in English;
- All students will develop high levels of proficiency in a L2. TWBI programs are considered additive bilingual programs for both groups of students because they afford all students the opportunity to maintain and develop oral and written skills in their first language while simultaneously acquiring oral and written skills in a second language;
- Academic performance for both groups of students will be at or above grade level, and the same academic standards and curriculum for other students in the district will also be maintained for students in TWBI programs; and

- All students in TWBI programs will demonstrate positive cross-cultural attitudes and behaviors.

Types of TWBI Programs. Despite the common characteristics among dual language programs, considerable variation exists in the languages used for instruction, the student population, and the time each language is used. TWBI programs also vary in how time is allocated for instruction in each language. The two basic models, the 90–10 model and the 50–50 model, vary in how they divide the time each language is used for instruction.

In the 90–10 model, the language other than English is used 90% of the time in early grades, and a gradually increasing proportion of instruction is done in English until sixth grade, when both languages are used equally in instruction. Many schools have adopted this model, placing an early emphasis on the language other than English to help compensate for the dominant power of English outside the school context. (Gómez, Freeman &Freeman, 2005)

In the 50–50 model, students learn in each language about half the time throughout the program. In many programs, all students learn to read in their primary language and then add the second language. Time for the two languages may be divided in various ways—half day and half day, alternate day, or even alternate week. This model is often used in areas with limited numbers of bilingual teachers. Teachers can team teach, and the bilingual teacher can provide the language other than English to one group in the morning and the other group in the afternoon (or on alternate days or weeks). This maximizes faculty language resources. (Gómez, Freeman &Freeman, 2005)

Effectiveness of Bilingual Education Programs

The effectiveness of bilingual education programs has been studied since they first were conceived thirty-five years ago. Several meta-analyses and research syntheses have attempted to

sum up what is known about the effectiveness of bilingual education programs (Greene, 1997; MacField, 2002; Rolstad, Mahoney, & Glass, 2005; Slavin & Chegun, 2005; Willig, 1985). Results of these five meta-analyses point in the direction of the advantage of bilingual education over all-English programs. The effectiveness of bilingual education continues to be highly debated in the U.S., despite a growing body of research that demonstrates its validity. After reviewing selected bilingual education studies, McField (2002) concluded that programs designed along principles hypothesized to underlie ideal bilingual programs were more effective. However, very few such comparisons were possible. Only one "strong" program and four "weak" programs could be analyzed in this way.

A study by Ramirez, Yuen, and Ramey (1991) was one of the first large longitudinal studies examining effectiveness of three types of transitional programs: early-exit, late-exit and structured English immersion. These researchers found no difference in levels of achievement or rates of growth in achievement in mathematics, English language, or English reading between students in structured English immersion and early-exit transitional programs after the end of third grade. However, students in both types of programs did better than students in the general population, and that late-exit transitional program students exhibited greater rates of growth than the general population. This study also found that all three types of programs resulted in student's increasing their skills in mathematics, English language, and reading as fast as or faster than students in the general population and that providing substantial instruction in the primary language did not impede developing language or reading skills in English.

Rossell and Baker (1996) used a categorical "vote counting" method to synthesize seventy-two (72) studies from 300 program evaluations nationwide comparing various forms of transitional and maintenance programs. They concluded that structured English immersion (also

called regular classroom English instruction) resulted in higher achievement gains than were found for all other forms of transitional and maintenance programs. Regrettably the review did not examine enrichment approaches to bilingual education.

Greene (1997) and Willig (1985) used meta-analysis to review bilingual education studies. Willig (1985) calculated effect sizes from sixteen studies and concluded that any type of bilingual education program is superior to no program. In other words, bilingual education is better than regular classroom instruction in English for language minority students. Using more rigorous methodological criteria, Greene (1997) estimated the effect sizes of eleven studies originally reviewed by Rossell and Baker (1996) and found contradictory results. He concluded that at least some native language instruction for language minority students was moderately beneficial compared to English-only approaches.

Two longitudinal studies by Thomas and Collier (1997, 2001) provide the most concrete information on the effectiveness of bilingual education. In their first study between 1982 and 1996, Thomas and Collier studied over 700,000 language minority students in five large urban and suburban school districts across the country. Two important findings from this study are highlighted below:

- Quality, long-term, enrichment bilingual programs that are well-implemented, give language minority students the best chance to succeed academically in English into the high school years.
- Many transitional and maintenance programs do not result in cognitively and academically prepared language minority students. Some transitional programs are no more successful than English-only programs in the long term.

In Thomas and Collier's second study (2001), over 200,000 students enrolled in five school districts between 1996 and 2001 were studied. Again, these researchers found that enrichment programs produce the highest achievement levels compared to other bilingual education programs. They also found:

- The more primary language grade-level schooling received, the higher students achieve in L2.
- It takes four to seven years of dual language schooling to begin outperforming other bilingually schooled students in all subject areas.
- Students with no primary language schooling are not able to reach grade-level performance in L2.
- Short-term, remedial programs do not close the achievement gap between language minority students and NES students.
- Students who receive at least 5 to 6 years of dual language instruction achieve parity in L2 by grade 5 of 6 and maintain that level of performance.

One additional large-scale study (Lindholm-Leary, 2001) found patterns of results on academic achievement for both native Spanish-speaking (NSS) and native English-speaking (NES) students that were similar to the Thomas and Collier studies.

Rigorous research studies on enrichment models of bilingual education are less common in the literature probably because these types of bilingual education programs have only begun to increase in popularity over the past decade. Most reports on these types of programs have been generated from school district sponsored program evaluations (see, for example, de Jong, 2002) and many are largely descriptive, correlational, or qualitative in design precluding causal links between programs and outcomes. The study by Hakuta, Bialystok, and Wiley (2003), for

example, confirmed Thomas and Collier's (1997) earlier judgment that younger students (particularly elementary school students) are most amenable to receiving the benefits of TWBI programs through an empirical analysis of census data using regression discontinuity modeling.

In recent years more quasi-experimental studies of TWBI programs have appeared in the literature. One outcome study of an enrichment bilingual education program that looked at achievement found that NES and NSS did not become equally bilingual and biliterate, but they did outperform their peers in their first and second language by the upper elementary grades (Freeman, 1998). Recent studies by Castillo (2001), Coy and Litherland (2000), Lucido and McEachern (2000), Sera (2000), and Stipek, Ryan, and Alarcón (2001) focused on academic achievement of early elementary students who were enrolled in TWBI programs and consistently reported achievement levels for NSS and NES in TWBI programs to be equal to or exceeding achievement levels of their peers in elementary schools that offered traditional bilingual education programs. More detailed descriptions of the results of these studies can be found in an excellent review by Howard, Sugarman, and Christian (2003). Similar findings were reported by Alanis (2000), Gilbert (2001), and Kortz (2002) in studies at the upper elementary school level.

Finally, a number of reports, texts, and journal articles have appeared recently describing characteristics of successful programs for language minority students. (c.f. August & Hakuta, 1997; August & Hakuta, 1998; Cloud, Genesee, & Hamayan, 2000; Doherty, Hilberg, Pinal, & Tharp, 2003; Escamilla, 2000; Lindholm-Leary, 2001; Senesac, 2002). Most of these reports echo the findings of Thomas and Collier's (1997; 2001) work, but also recommend more generic features such as assessment and accountability components, connections with parents and the community, etc.

CHAPTER 3: METHOD

The basis for conducting this study of a two-way bilingual immersion education program was to investigate the effects of the program on reading, writing, and math achievement in this study population, and to promote the use of methodologically sound comparative research in the bilingual education field. A detail description of all aspects of the design and procedures used in this study are presented in this chapter.

Research Design and Rationale

This is a longitudinal quasi-experimental study. The study design was an ex post facto, non-randomized, matched-pairs design. A multi-level matching procedure was used to match students from the two-way bilingual immersion (TWBI) elementary school with a control group beginning in third grade. Data on CSAP reading, writing, and math achievement was collected at sixth, seventh, eight, nine, and tenth grades. Data on ACT English, writing, mathematics and science was collected at eleventh grade.

The advantages of longitudinal studies over other study designs, such as cross-sectional studies, are well documented. Longitudinal data involve repeated measures of the same subjects over time, while cross-sectional data involve measures at one time only. Thus, cross-sectional research can only measure the prevalence of a factor of interest at a certain point in time, while longitudinal research measures prevalence at several points in time, and can provide some information on causation, and change (Menard, 2007).

Longitudinal studies enable factors of interest to be examined for stability and continuity, and can identify developments over time. Longitudinal studies also allow researchers to differentiate between change over time in aggregate (group) data. While cross-sectional data only allow investigation of differences between individuals, a longitudinal study can examine

change within individuals as well as variation between them. Repeated measures allow for the detection of change in individuals or their environments from one data point to the next (Black, 1991).

A matched pairs design involves using different but similar participants in each condition. If there are any important characteristics that might affect performance, researchers will try to match participants on those characteristics in each condition. Two advantages of this design are: participant variables are kept more constant between conditions if matched; and more sophisticated statistical tests can be used because of less variation between conditions. There are also disadvantages associated with the matched pair design:

- Participant variables can never be perfectly matched in every way; they may still affect results.
- Matching participants is time consuming and can be difficult.
- In preparing for the study more people are required in order to ensure good matches.
- It may be difficult to identify appropriate criteria for matching (Love, 2005).
- Attrition in one member of the pair usually means both participants would be omitted from the analysis.

Pilot Study

A pilot study was conducted between 2002 and 2005 and published in the Middle Grades Research Journal (MGRJ) (see Cobb, Vega & Kronauge, 2006). Later, this article was selected to be part of collection of studies published in the book "Middle Grades Research: Exemplary Studies Linking Theory to Practice". This book was published in 2009 and presents a thoroughly scrutinized group of studies focusing on middle grades education issues. As a collection, the ten studies featured in this book are the most highly peer-reviewed manuscripts examined, between

August 2006 and December 2008, by members of the MGRJ Review Board -- each having undergone careful "blinded" examination by three or more experts in the sub-specialty area addressed by the research study conducted. This article has been cited by 9 scholarly works, including books, journal articles, and dissertations (e.g., De Jong & Bearse, 2011; García & Jensen, 2010; Shneyderman & Abella, 2009).

The pilot study compared four cohorts of native English-speaking students (NES) and native Spanish-speaking students (NSS) from the TWBI elementary school with their matched pairs from comparable programs within the school district. Students were matched in third grade and comparisons were made in sixth and seventh grades. Only 31 pairs of students (10 NSS and 21 NES) were analyzed in this study.

The fundamental hypothesis was that the process of receiving instruction in two languages (English and Spanish) throughout elementary school (i.e., attendance at a dual language school) would help the at-risk NSS students in the school and not have a negative effect on the NES students in the school in core academic areas (reading, mathematics, writing), and that this beneficial effect would carry through the first year of junior high school in which instruction was delivered through a traditional English-only model. The pilot study demonstrated that dual language schooling, when implemented properly by schools, must be considered at least equally as effective in core academic achievement areas as traditional elementary schooling, and is probably more effective in the long term.

The present study is using the same design and analysis that was used in the pilot study but it is adding 7 new cohorts, another outcome variable (ACT scores) and comparisons in 8^{th}, 9^{th}, 10^{th}, and 11^{th} grades. A detail description of the procedures follows.

Data Collection

Data were collected through school district records on NES and NSS students from the two-way bilingual elementary school and their matched pairs who were students from comparable programs within the school district. Eleven annual cohorts of NES and NSS students from the experimental school were matched on a student-by-student basis on seven variables – cohort year, student's primary language, years of enrollment in the program, ethnicity, gender, socioeconomic status, and 3rd grade performance – with comparable students from within the school district. Comparable students mean that they had very similar characteristics to the ones in to the treatment group in regard to the matching variables. In some variables, all the treatment students were matched perfectly (i.e. cohort year and primary language) while in other, the matched student had similar characteristics. A detail description of the comparability of the samples is presented in the section "Matching Procedure" in this chapter.

The eleven cohorts included in the sample consisted of students who were enrolled in third grade in any of the following academic years: 1996-97, 1997-98, 1998-99, 1999-00; 2000-01; 2001-02; 2002-03; 2003-04; 2004-05; 2005-06; or 2006-07.

All experimental and control students must have been enrolled in their schools (experimental or control) for a minimum of four years to be included in this study. These eleven cohorts of 3rd graders were then tracked to middle and high school and measured on their reading, writing, and math achievement scores in sixth, seventh, eighth, ninth, and tenth, grades. ACT scores were collected in eleventh grade.

Population, Sample, and Sampling Design

The study population was selected from a school district in Northern Colorado. The district consisted of 27 elementary schools, 8 middle schools, and 4 high schools in a city with a

population of approximately 120,000 and total enrollment of approximately 25,000. Approximately eight percent of the district's student population is considered as English language learner (ELL) (Colorado Department of Education, 2002; Escamilla, 2000). The district offers two types of programs for ELL students: TWBI (enrichment) and English as a Second Language (ESL) (transition).

Sampling procedure. Inclusionary and exclusionary criteria were defined to create four different samples: NES students enrolled in the TWBI program, NSS students enrolled in the TWBI program, NES students enrolled in the standard elementary school program, and NSS students enrolled in a comprehensive ESL program housed in a traditional elementary school. Exclusionary criteria were the same for all four samples: (a) absence of primary language information, (b) qualification for special education services, and (c) enrollment in their specific program/school less than four years. The TWBI program (experimental) samples were created using the following inclusionary criteria: (1) the student must have been enrolled in third grade in 1996-97, 1997-98, 1998-99, 1999-00; 2000-01; 2001-02; 2002-03; 2003-04; 2004-05; 2005-06; or 2006-07; (2) there must have been accurate records on student ethnicity, gender, socioeconomic status, and a valid third grade high stakes reading achievement score, and (3) the student must have been enrolled in the program for at least four years (i.e., from first to fourth grade, from second to fifth grade, or from third to six grade). The only difference in the process of creating each of the two TWBI groups (NES and NSS students) was that each NES student must have had a valid third-grade reading achievement score, while each NSS student must have had a valid third-grade English-proficiency score from the school district ELA's (English Language Acquisition) office. This differential matching criterion existed because NSS students from the experimental group did not get tested for reading ability in English until later in the

bilingual program; therefore the same measure was not available for them. The control program samples were created through the matching procedure described below.

Matching procedure. A multi-level matched pairs selection procedure was used to select the comparable NES and NSS students. The matching procedures were different depending on the student's native language. The available pool of non-TWBI NES students from which to match the TWBI NES students was significantly greater than the available pool of non-TWBI NSS students from which to match the TWBI NSS students. In addition, district records on native Spanish-speaking students were limited.

To select each non-TWBI NES student to pair with each TWBI NES student, the following procedure was used. First, the two schools that most closely resembled the size and the demographic characteristics of those of the English-speaking population attending the TWBI school were selected. Second, a pool of students from these schools that met the inclusion and exclusion criteria previously described for English-speaking students was created. Third, TWBI students were matched with students from the pool on the following variables: native language, year student was enrolled in 3rd grade, number of years enrolled in the program, ethnicity, gender, eligibility for meal benefits, and 3rd grade reading achievement test score. This procedure was done when the student was in 6th grade, so it was done retrospectively. Students from the treatment group kept the same match student they had in 6th grade across all the years they were followed. Therefore, it is expected that the higher the grade, the less matched pairs remain in the study.

The procedure used to match NSS students was different for reasons previously explained. Instead of creating a pool of students from schools that were similar to the TWBI school, a district-wide pool of NSS students was created. Students in this pool were then

screened to meet the inclusion and exclusion criteria previously described for NSS students. Finally, they were matched with TWBI NSS students on the following variables: native language, year student was enrolled in 3rd grade, number of years enrolled in the program, ethnicity, gender, eligibility for meal benefits, and English proficiency score. Matching on these

Table 2

Distributions of Study Sample on Matching Variables

	Treatment Students		Control Students	
	n	%	*n*	%
Cohort Year				
1996-1997	19	5.4	19	5.4
1997-1998	19	5.4	19	5.4
1998-1999	18	5.2	18	5.2
1999-2000	27	7.7	27	7.7
2000-2001	34	9.7	34	9.7
2001-2002	34	9.7	34	9.7
2002-2003	37	10.6	37	10.6
2003-2004	43	12.3	43	12.3
2004-2005	34	9.7	34	9.7
2005-2006	36	10.3	36	10.3
2006-2007	48	13.8	48	13.8
Primary Language				
Spanish	184	52.7	184	52.7
English	165	47.3	165	47.3
Ethnicity				
Asian or Pacific Islander	4	1.1	3	0.9
African American	5	1.4	3	0.9
White (not Hispanic)	112	32.1	115	33.0
Hispanic	228	65.3	228	65.3
Socio-economic status				
Low	198	56.7	195	55.9
Not-Low	151	43.3	154	44.1
Years Enrolled in Program				
3 or less	0	0.0	11	3.2
4 to 6	349	100.0	338	96.8
Gender				
Male	163	46.7	163	46.7
Female	186	53.3	186	53.3

factors was not perfect in all cases. A summary of descriptive information on the matched pairs is presented in Table 2. A perfect match for all treatment students was achieved in two variables: cohort year and primary language, where 100% of the pairs have the same values. Only 3 participants of the 349 included in the study were not matched with a pair of the same ethnicity. The number of mismatches in gender was 16 for native English-speakers and 10 for native Spanish-speakers. There were only 3 native Spanish-speakers and 16 native English-speakers that were not matched with a pair of the same socio-economic status (SES). These 16 English-speaking participants from the treatment group were classified as Low SES and they were matched with a pair of not-Low SES. In other words, in the absence of an equal match for the experimental student, one with a better condition was chosen from the match sample.

The average number of years of enrollment for the treatment group was 5.61 with a standard deviation of .62, and 5.50 years with .85 for the control group.

NES were matched on their 3rd grade CSAP reading score. The average score for the treatment group was 597.64 with a standard deviation of 69.92 and for the control group was 600.23 with a standard deviation of 69.94. Conversely, NSS were matched on their third-grade English-proficiency score assigned by the school district ELA's (English Language Acquisition) office. The ELA´s office classified students in one of three levels of English proficiency. Out of the 185 NSS pairs, 168 (90.8%) had a match with the same level of English proficiency. The average classification level for the treatment group was 1.47 with a standard deviation of .82 and for the control group was 1.49 with a standard deviation of .83.

Final Samples and Attrition

The study population consisted of all students registered in third grade between the academic years of 1996 and 2007 in a mid-sized school district in the western region of the

country. Of these third grade students, every student at the study school was initially eligible to participate, as well as a matched student who was paired with the treatment student but enrolled in a different school in the same school district.

Inclusion criteria for students from the treatment school to participate in the study were: enrolled at Treatment program for at least 4 years, not receiving special education services, and availability of records. Out of the 466 eligible students, 15 were excluded because they were receiving special education services and 102 students were excluded because they were enrolled for less than 4 years in the treatment school. No student was excluded for unavailability of demographic records. As shown in Table 3, almost 75% of all the students registered in third grade were included in the study.

Table 3

Treatment Students Included in the Study as a Percentage of Students Registered in Third Grade in the Treatment Program by Academic Year

Academic Year	Registered in 3rd grade	Included in study	
	N	*n*	%
96-97	30	19	63.3%
97-98	30	19	63.3%
98-99	29	18	62.1%
99-00	50	27	54.0%
00-01	48	34	70.8%
01-02	43	34	79.1%
02-03	43	37	86.0%
03-04	49	43	87.8%
04-05	45	34	75.6%
05-06	48	36	75.0%
06-07	51	48	94.1%
Total	466	349	74.9%

Attrition rate at this stage of the study was higher for Spanish speaking students. As shown in Table 4, a total of 82.5% of the native English-speaking students were included in the study, versus 69.2% of the native Spanish-speaking students. It is probable that this phenomenon

was the result of the characteristics of the Spanish speaking population who attended the treatment program, comprised mainly of low-income immigrant families who tend to move more frequently.

Table 4

Treatment Students Included in the Study as a Percentage of Students Registered in Third Grade in the Treatment Program by Language and by Academic Year

Academic Year	Language of student	Registered in 3rd grade	Included in study	
		N	*n*	%
96-97	English	14	12	85.7%
	Spanish	16	7	43.8%
97-98	English	13	11	84.6%
	Spanish	17	8	47.1%
98-99	English	18	11	61.1%
	Spanish	11	7	63.6%
99-00	English	25	16	64.0%
	Spanish	25	11	44.0%
00-01	English	17	11	64.7%
	Spanish	31	23	74.2%
01-02	English	21	16	76.2%
	Spanish	22	18	81.8%
02-03	English	18	15	83.3%
	Spanish	25	22	88.0%
03-04	English	21	20	95.2%
	Spanish	28	23	82.1%
04-05	English	17	16	94.1%
	Spanish	28	18	64.3%
05-06	English	17	17	100.0%
	Spanish	31	18	58.1%
06-07	English	19	19	100.0%
	Spanish	32	29	90.6%
All cohorts	English	200	165	82.5%
	Spanish	266	184	69.2%

To study attrition, we examined all treatment students with a valid 6th grade CSAP Reading score who were matched with a control student (see Table 5). Students from cohort 1996-1997 did not have a valid 6th grade CSAP reading score because there was no CSAP testing

for 6th grade during the year that cohort was in 6th grade (1999-2000 academic year). These students were kept in the study because they met criteria for inclusion and they had valid records for 7th, 8th, 9th, 10th CSAP scores and ACT scores.

Table 5

Treatment Students with Valid 6th grade CSAP Reading Score and Treatment Students Still Matched to Student with Valid 6th Grade CSAP Reading Score as a Percentage of Students Registered in Third Grade in the Treatment Program by Academic Year

Academic Year	Registered in 3rd grade	With valid 6th grade CSAP reading score		Matched to a student with a valid 6th grade CSAP reading score	
	N	*n*	%	*n*	%
96-97	30	0	0%	0	0%
97-98	30	16	53%	16	53%
98-99	29	18	62%	18	62%
99-00	50	26	52%	26	52%
00-01	48	34	71%	34	71%
01-02	43	34	79%	34	79%
02-03	43	36	84%	36	84%
03-04	49	40	82%	40	82%
04-05	45	34	76%	34	76%
05-06	48	36	75%	36	75%
06-07	51	48	94%	48	94%
Total	466	322	69%	322	69%

It was expected that as students advanced in their academic path, fewer of them will remain in the study. Conversely, it was not expected that students from the control group had a higher attrition rate that students from the treatment group. For example, 43 students were registered in 3rd grade in the treatment program during the 2001-2002 academic. Out of those 43, only 34 students were in the school district and had a valid CSAP reading score when they were in 6th grade. Those 34 students were matched with students from comparable programs following the matching procedure described earlier. Consequently, there were also 34 students in the control group with valid CSAP reading scores (see table 5). When these two groups of

students reached 10th grade, there were 28 students from the treatment group who remained in the school district and had a valid 10th CSAP reading score, however there were only 20 students from the control group that did (see table 6). This is evidence that attrition rate was higher for the control group than for the treatment group.

Out of the 273 students from the treatment program registered in 3rd grade during the period of the study, only 162 (59%) where in the school district and had a valid CSAP Reading score when they got to 10th grade. However, only 48% of the students from the control group did (see Table 6).

Table 6

Attrition Rate at 10th grade. Treatment Students with Valid 10th grade CSAP Reading Score and Treatment Students Still Matched to Student with Valid 10th Grade CSAP Reading Score as a Percentage of Students Registered in Third Grade in the Treatment Program by Academic Year

Academic Year	Registered in 3rd grade	With valid 10th grade CSAP reading score		Matched to a student with a valid 10th grade CSAP reading score	
	N	*n*	%	*n*	%
96-97	30	17	57%	14	47%
97-98	30	14	47%	12	40%
98-99	29	17	59%	10	34%
99-00	50	24	48%	23	46%
00-01	48	33	69%	23	48%
01-02	43	28	65%	20	47%
02-03	43	29	67%	28	65%
Total	273	162	59%	130	48%

This trend continued into 11th grade, where the percentages are even lower. For example, out of the 43 students were registered in 3rd grade in the treatment program during the 2001-2002 academic, 21 were in the school district and had a valid ACT English score when they were in 11th grade, but only 13 from the control group that did (see table 7).

Table 7

Attrition Rate at 11th grade. Treatment Students with Valid 11th grade ACT English Score and Treatment Students Matched to Student with Valid 11th Grade ACT English Score as a Percentage of Students Registered in Third Grade in the Treatment Program by Academic Year and by Language

Academic Year	Language of student	Registered in 3rd grade	With valid ACT English score		Matched to a student with a valid ACT English score	
		N	*n*	%	*n*	%
96-97						
	English	14	11	79%	10	71%
	Spanish	16	6	38%	2	13%
	All	30	17	57%	12	40%
97-98						
	English	13	10	77%	8	62%
	Spanish	17	6	35%	1	6%
	All	30	15	50%	9	30%
98-99						
	English	18	8	44%	6	33%
	Spanish	11	6	55%	2	18%
	All	29	14	48%	8	28%
99-00						
	English	25	13	52%	8	32%
	Spanish	25	8	32%	4	16%
	All	50	21	42%	12	24%
00-01						
	English	17	8	47%	6	35%
	Spanish	31	13	42%	3	10%
	All	48	21	44%	9	19%
01-02						
	English	21	11	52%	9	43%
	Spanish	22	10	45%	4	18%
	All	43	21	49%	13	30%
Total						
	English	122	61	50%	47	39%
	Spanish	108	48	44%	16	15%
	All	230	109	47%	63	27%

Out of the 230 students from the treatment program registered in 3rd grade during the period of the study, only 109 (47%) remained in the school district and had a valid ACT English

score when they got to 11th grade. However, only 27% of the students from the control group did (see Table 7).

Results from the attrition analysis by language were strikingly unexpected, especially for upper grades. For example, out of the 122 NES students from the treatment program registered in 3rd grade during the period of the study, only 61 (50%) remained in the school district and had a valid ACT English score when they got to 11th grade. Out of the 108 NSS students from the treatment program registered in 3rd grade during the period of the study, only 48 (44%) remained in the school district and had a valid ACT English score when they got to 11th grade. However, NSS students from the matched sample were present in lower numbers at 11th grade, only 16 (15%), compared to 47 (39%) for NES (see Table 7).

Instrumentation

Four instruments were used in this study. The Northwest Evaluation Association's (NWEA) Achievement Levels Test and the Language Assessment Scales (LAS), were used for matching pairs of students. Other two instruments, the CSAP reading, writing, and math subtests, and the ACT, were used to compare student´s performance. Following is a description of each of these instruments.

NWEA achievement levels test. The NWEA Levels Test is a norm referenced achievement test for elementary and middle school students that is used in many school districts in the US. The NWEA maintains large item banks in four content areas: language, math, reading, and science. Rasch IRT is used to calibrate the tests. Studies have found the tests to have high reliability (greater than .90). The most recent norming study was done in 1995.

Language assessment scales. The LAS measures English language skill in reading and writing in language-minority students across seven content areas. Those areas are vocabulary,

fluency, reading for information, mechanics and usage, sentence completion, sentence writing, and essay writing. The LAS is used as a screening tool to provide placement and classification information for language-minority students. The LAS test is available in two alternate forms and has three levels for Grades 2 through high school. Each test contains 45 multiple choice questions, one story, five open-ended questions, and over 5 graphic prompts to elicit written responses.

CSAP test. The CSAP test was designed to determine what levels students throughout Colorado meet the Colorado Model Content Standards in reading, writing, math, and science. The CSAP is a criterion-referenced test that was developed by the Colorado Department of Education, a test developer, and teachers and curriculum specialists from around the state of Colorado. The CSAP test was first administered during 1996-1997 and now given in grades three through ten. Question items on the CSAP are either multiple choice or constructed response. Students taking the test receive a score and performance level. The CSAP test has been found to have high internal consistency across all content areas and grade levels (Cronbach's alpha range between .88 and .93) (Colorado Department of Education, 2002).

ACT. The ACT is a national college admission and placement examination test, first administered in 1959. It is curriculum-based and not an aptitude or an IQ test. Instead, the questions on the ACT are directly related to what students have learned in high school courses in English, mathematics, and science. The ACT questions cover the following subject areas: English with 75 questions, math with 60, reading with 40, science with 40, and writing which consist of a 30- minutes essay test. The minimum possible score is 0 and the maximum is 36 (ACT, 2007). In 2011 the ACT was taken by 49% of U.S. high school graduates (ACT, 2012a). The Composite Score is the average of the four test scores, rounded to the nearest whole number.

The national average ACT composite score for 2012 was 21.1. During the same year, the national average ACT score for English was 20.5, for reading was 21.3, for mathematics was 21.1, and for science 20.9. Test scores remained essentially the same between 2008 and 2012 even though about 17% more high school students took the ACT over this period and the tested population of students became more diverse (ACT, 2012b).

Planned Analysis

A total of eleven cohorts of students were grouped to perform analysis on academic achievement of sixth graders. A total of six cohorts of students were grouped to perform analysis on academic achievement of eleventh graders (see Figure 2), which shows that five of the eleven cohorts had not reached 11th grade by the time the data were compiled in 2009-2010. A total of seven cohorts of NES and NSS students and their match pairs were compared on reading, writing, and math achievement scores in sixth, eight, ninth, and tenth grade.

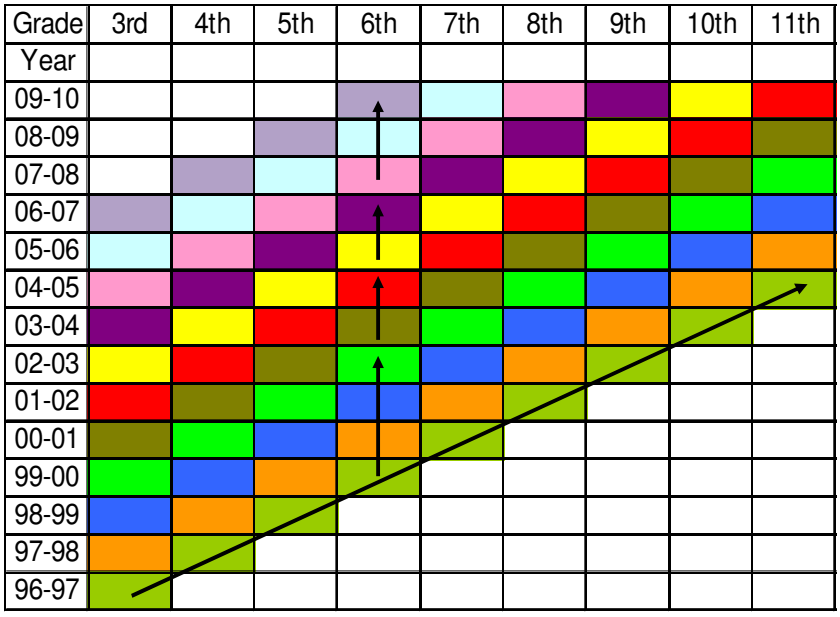

Figure 2. Data collection chart across time and grade level.

. Means, standard deviations, sample sizes, and effect sizes for each group in each cohort were calculated. It has been suggested that an effect size of .20 represents a small impact of a treatment, while .50 represents a modest impact and .80 represents a large impact (Cohen, 1988).

The between groups independent variable for this study was "language" with two levels – Spanish and English. There were two within-subjects independent variables. Because this was a matched pairs design, the primary independent variable "program" was analyzed as a within-subjects variable with two levels – TWBI or treatment program and traditional program. The second within subjects variable was "change over time" with five levels – 6th grade, 7th grade, 8th grade, 9th grade, and 10th grade. The dependent variables for this study were reading, writing, and mathematics CSAP scaled scores for both levels of the within subject independent variable. Hence this design is classified as a 2 x 2 x 5 mixed design with repeated measures on the second and third factors. ACT was also a dependent variable for the 11th grade students.

Analysis of variance was used to evaluate significant statistical differences between the program groups´ academic performance. The Statistical Package for the Social Sciences (SPSS) version 19.0 was used for the analyses of the data. Effect sizes were calculated using standard formulae for the Hedges g statistic.

To address the first hypothesis, a set of three 2 x 2 x 5 mixed ANOVA's (one each in reading, writing, and mathematics respectively) with repeated measures on the second and third factor will be run. The three independent variables in each of the 2 x 2 x 5 mixed ANOVA´s are, in sequence: primary language (English or Spanish) – a between-groups factor, type of schooling or program (treatment program or control program) – a within-groups factor, and grade level (6th, 7th, 8th, 9th or 10th) – a repeated -measures factor. To interpret the analyses, we will look for statistically significant main effects on the type of program. We will also look for the absence of

statistically significant interactions with program, on the grade level factor in the mixed ANOVA´s

A separate analysis will be conducted on 6th grade CSAP data because this is the last grade of the elementary school. Therefore, another set of three 2 x 2 mixed ANOVA´s for CSAP 6th data (one each in reading, writing, and mathematics respectively) will be run. The two independent variables in each of the 2 x 2 ANOVA´s will be, in sequence: primary language (English or Spanish) – a between-groups factor, and type of schooling or program (treatment program or control program) – a within-groups factor.

For the second hypothesis, we will run three (reading, writing, and math) 2x5 Repeated Measures ANOVAs with NSS only. The two factors will be both within-subjects factors (type of program and time). Then, language will not be a factor because we will be using only NSS.

For hypothesis number three, we will run three (reading, writing, and math) 2x5 Repeated Measures ANOVAs with NES only. The two factors will be within-subjects factors (type of program and time). Again, language will not be a factor because we will be using only NES.

To address the fourth hypothesis, a set of four 2 x 2 mixed ANOVA´s for ACT 11th grade data (one each in English, reading, mathematics, and science respectively) will be perform. The two independent variables in each of the 2 x 2 mixed ANOVA´s will be, in sequence: primary language (English or Spanish) – a between-groups factor, and type of schooling or program (treatment program or control program) – a within-groups factor.

The evaluation of hypothesis five will be done by performing four (ACT English, reading, science, and math) paired *t* tests with NSS only. And finally, hypothesis six will be tested by running four (ACT English, reading, science, and math) paired *t* tests with NES only.

CHAPTER 4: RESULTS

Introduction

The purpose of this study was to investigate the effects of a well implemented TWBI (two way bilingual immersion) education program on student achievement. More precisely, this study attempted to demonstrate that the process of receiving instruction in two languages (English and Spanish) throughout elementary school (attendance at a TWBI school) would help the Native Spanish-speaking students (NSS) in the school and not have a negative effect on the Native English-speaking students (NES) in the school in core academic areas (reading, writing, and mathematics). Furthermore, this beneficial effect would carry through junior high and high school in which instruction was delivered through a "business as usual" English-only model. Analysis of variance was used to evaluate significant statistical differences between groups in academic performance. The Statistical Package for the Social Sciences (SPSS) was used for the analyses of the data. Effect sizes were calculated using standard formulae for the Hedges g statistic. We calculated these effect sizes for all comparisons of interest regardless of statistical significance in order to help understand and interpret patterns of findings. Results in this section are organized and presented by research hypothesis. However, the six hypotheses can be grouped in two sets of three. The first set refers to the academic performance in grades 6 through 10 of the experimental group as a whole versus the control group, as well of comparisons native English-speaking and native Spanish-speaking students against their matched pairs. The second set of hypotheses refers to performance on ACT scores.

The first set of three hypotheses was:

1. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in reading, mathematics, and writing at

the end of 6th, 7th, 8th, 9th, and 10th grades than will a matched pair´s control group as measured by the Colorado Student Assessment Program test battery.

2. Those native Spanish-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in each of these achievement domains and each of these five years than a matched control group.
3. Those native English-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in each of these achievement domains and each of these five years as a matched control group.

First, an analysis was conducted on the academic performance of students at the last grade of the elementary school (6th grade). The rationale for this analysis was to establish if differences in academic performance were present at this point in time. A set of three 2 x 2 mixed ANOVA´s for CSAP 6th grade data (one each in reading, writing, and mathematics respectively) were conducted. The two independent variables in each of the 2 x 2 ANOVA´s were, in sequence: primary language (English or Spanish) – a between-groups factor, and type of schooling (treatment program or control program) – a within-subjects factor.

To address the first hypothesis, a set of three 2 x 2 x 5 mixed ANOVA's (one each in reading, writing, and mathematics respectively) with repeated measures on the second and third factor was performed. The three independent variables in each of the 2 x 2 x 5 mixed ANOVA´s were, in sequence: primary language (English or Spanish) – a between- groups factor, type of schooling (treatment program or control program) – a repeated-measures factor, and grade level (6th, 7th, 8th, 9th or 10th) – a repeated-measures factor.

For the second hypothesis, we ran three (reading, writing, and math) 2 x 5 repeated-measures ANOVAs with NSS only. The two factors both were within-subjects factors (type of program and time).

For hypothesis number three, we ran three (reading, writing, and math) 2 x 5 repeated-measures ANOVAs with NES only. The two factors both were within-subjects factors (type of program and time).

Interpretation of the results was done by looking for consistent patterns across all ANOVA's, and specifically for statistically significant main effects on the type of program.

The second set of three hypotheses was:

4. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group as measured by the American College Testing test battery.
5. Those native Spanish-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group.
6. Those native English-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in English, reading, mathematics, and science in 11th grade as a matched control group.

To address these three hypotheses the following analyses were performed. To address the fourth hypothesis, a set of four 2 x 2 mixed ANOVA´s for ACT 11th grade data (one each in English, reading, mathematics, and science respectively) were performed. The two independent variables in each of the 2 x 2 mixed ANOVA´s were, in sequence: primary language (English or

Spanish) – a between-groups factor, and type of schooling (treatment program or control program) – a within-subjects factor.

The evaluation of hypothesis five was done by performing four (ACT English, reading, science, and math) paired *t* tests with native Spanish-speaking students only.

And finally, hypothesis six was tested by running four (ACT English, reading, science, and math) paired *t* tests with native English-speaking students only.

An examination for conformity to the assumptions underlying each mixed ANOVA was conducted. The assumptions for mixed ANOVAs "include independence of observations (unless the dependent data comprise the within-subjects factor), normality, and homogeneity of variances…, known as sphericity" (Leech, Barrett & Morgan, 2005, p. 147). Sphericity was not an issued in many of these analyses because there were only two levels in the program type, a within-subjects factor. For mixed ANOVAs that involved a factor with more than two levels (i.e., grade level), Greenhouse-Geisser-corrected degrees of freedom were employed in all *F*-tests involving main effects and interactions of these repeated measures.

TWBI Students Performance at the End of Elementary School

The CSAP means, standard deviations, and sample sizes for each 6th grade group are presented in the Table 8. The number of students included here (i.e., 171 matched pairs of native Spanish speakers and 151 native English speakers for reading) is significantly higher than the number of students included in the later analyses. This is due to the fact that all matched students with valid results for 6th grade were included, but some students did not have data for 6th through 10th grades.

As can be observed in Table 8, all CSAP mean scores are somewhat higher for students in the Treatment Program than in the Control Program, regardless of language group. The

biggest difference, more than 25 points, was located in the math subject within the Spanish-speaking students group.

Table 8

CSAP 6th Grade Scaled Scores Broken out by Language Group and Intervention

	Reading			Writing			Math		
	n	Mean	Std. Dev.	*n*	Mean	Std. Dev.	*n*	Mean	Std. Dev.
Spanish Speakers									
Treatment	171	601.42	54.78	166	511.04	45.89	166	524.96	66.14
Control	171	592.70	52.73	166	501.16	45.87	166	496.39	58.73
English Speakers									
Treatment	151	682.39	47.73	135	582.36	51.06	141	603.59	55.27
Control	151	667.20	52.22	135	566.21	47.65	141	587.12	67.53

For the 6th grade data, a 2 x 2 mixed ANOVA of CSAP reading scores yielded a statistically significant type of program effect, $F(1, 320) = 13.35$, $p < .001$; a non-significant type of program by language interaction effect, $F(1, 320) = .98$, $p = .323$; and a significant primary language effect, $F(1, 320) = 261.81$, $p < .001$ (see Table 9). Both main effects were significant,

Table 9

Mixed ANOVA Results for 6th Grade CSAP Reading Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	969,083.89	1.00	969,083.89	261.81	<.001
Error 1	1,184,458.48	320	3,701.43		
	Within subjects				
Type of Program	22,924.27	1.00	22,924.27	13.35	<.001
Type of Program x Primary Language	1,679.83	1.00	1,679.83	.98	.323
Error 2	549,324.98	320.00	1,716.64		

however, only the type of program effect was of interest for this study.

Post-hoc comparisons of the means were not performed because there were fewer than three groups in each of the variables, and the direct observation of means allow one to conclude that performance of students from the treatment program was higher. A 2 x 2 mixed ANOVA of CSAP writing scores yielded the same results, a statistically significant type of program effect, $F(1, 299) = 19.26$, $p < .001$; a non-significant type of program by language effect, $F(1, 299) = 1.12$, $p = .292$; and a significant primary language effect, $F(1, 299) = 216.36$, $p < .001$ (see Table 10). Both main effects were significant; however, only the type of program effect was of interest for this study.

Table 10

Mixed ANOVA Results for 6th Grade CSAP Writing Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	692,359.17	1.00	692,359.17	216.36	<.001
Error 1	956,24.46	299	3,200.08		
	Within subjects				
Type of Program	25,203.95	1.00	25,203.95	19.26	<.001
Type of Program x Primary Language	1,459.36	1.00	1,459.36	1.12	.292
Error 2	391,226.96	299.00	1,308.45		

The third 2 x 2 mixed ANOVA of CSAP on math scores yielded also the same results, a statistically significant type of program effect, $F(1, 305) = 31.00$, $p < .001$; a non-significant type of program by language interaction effect, $F(1, 305) = 2.24$, $p = .136$; and a significant primary language effect, $F(1, 299) = 208.97$, $p < .001$ (see Table 11).

This result of the main effects and the non-significant interactions indicate that students who attended the treatment program had significantly higher CSAP reading, writing and math scores at the end of their elementary school when compared with their matched pairs, regardless of their native language.

Table 11

Mixed ANOVA Results for 6th Grade CSAP Math Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,093,479.14	1.00	1,093,479.14	208.97	<.001
Error 1	1,596,000.25	305	5,232.79		
	Within subjects				
Type of Program	77.332,64	1.00	77,332.64	31.00	<.001
Type of Program x Primary Language	5,585.10	1.00	5,585.10	2.24	.136
Error 2	760,918.87	305.00	2,494.82		

TWBI Students Performance during Junior High and High School

Longitudinal data was analyzed using mixed ANOVA tests. With respect to our first and second research hypotheses, we expected to observe a consistent main effect on program type across all three outcome domains, indicating the strength and breadth of the intervention's effect in general, and specifically for native Spanish-speaking students. With respect to the third research hypothesis, we expected to see a null program type main effect, and a null program type x time interaction effect across all three outcome domains, indicating no adverse effect on native English-speaking children associated with receiving half of their instruction in Spanish

throughout their elementary school years. We were not interested in either the language or time main effects as these would be expected independent of the treatment's effects.

Hypothesis 1. Only pairs of students with valid records for all grade levels in each subject area were included in these analyses. Data were initially examined using both Mauchly´s Test of Sphericity and Levene´s Test of Equality of Error Variances. When the assumption of sphericity was violated, Greehhouse-Geisser correction was used in the analysis (Leech, Barrett & Morgan, 2005, p. 151). The assumption of equality of variances was not violated. Sample

Table 12

CSAP Reading, Writing, and Math Average Scaled Scores Broken out by Grade and Type of Intervention for Native Spanish-Speaking Students with Effect Sizes for Simple Effects

	Treatment			Control			Cohen's
	n	Mean	Std. Dev.	*n*	Mean	Std. Dev.	*d*
Sixth Grade							
Reading	42	605.00	42.00	42	590.31	49.00	0.32
Writing	40	512.83	45.84	40	506.63	40.66	0.14
Math	44	514.23	58.36	44	489.39	58.66	0.42
Seventh Grade							
Reading	42	611.12	55.08	42	594.64	46.41	0.32
Writing	40	519.33	57.88	40	512.45	48.32	0.13
Math	44	518.80	61.08	44	489.36	67.32	0.46
Eighth Grade							
Reading	42	624.62	39.93	42	603.79	61.52	0.41
Writing	40	525.38	58.83	40	514.43	49.62	0.20
Math	44	535.57	51.97	44	503.48	68.11	0.53
Ninth Grade							
Reading	42	645.45	33.26	42	626.24	40.92	0.52
Writing	40	544.45	65.79	40	528.30	51.64	0.28
Math	44	547.02	66.45	44	520.27	79.94	0.37
Tenth Grade							
Reading	42	659.43	46.41	42	636.86	49.07	0.47
Writing	40	539.28	70.91	40	518.10	62.90	0.32
Math	44	555.89	63.56	44	522.52	71.24	0.50

sizes, means, standard deviations, and effect sizes are presented in the Table 12 for native Spanish-speaking students and in Table 13 for native English-speaking students.

Table 13

CSAP Reading, Writing, and Math Average Scaled Scores Broken out by Grade and Type of Intervention for Native English-Speaking Students with Effect Sizes for Simple Effects

	Treatment			Control			Cohen's
	n	Mean	Std. Dev.	*n*	Mean	Std. Dev.	*d*
Sixth Grade							
Reading	49	685.53	62.13	49	662.12	54.35	0.40
Writing	43	590.72	59.74	43	571.28	42.45	0.38
Math	45	606.56	72.42	45	588.04	68.98	0.26
Seventh Grade							
Reading	49	688.14	49.27	49	671.59	52.88	0.32
Writing	43	608.30	61.22	43	598.84	53.24	0.17
Math	45	599.47	55.65	45	597.53	62.01	0.03
Eighth Grade							
Reading	49	703.41	48.79	49	685.39	49.79	0.37
Writing	43	613.93	66.59	43	611.26	63.30	0.04
Math	45	611.91	58.24	45	608.36	58.71	0.06
Ninth Grade							
Reading	49	710.59	40.62	49	702.79	45.47	0.18
Writing	43	639.40	73.43	43	618.88	64.53	0.28
Math	45	630.13	54.30	45	626.42	55.00	0.07
Tenth Grade							
Reading	49	721.39	38.06	49	709.63	52.14	0.26
Writing	43	648.60	87.62	43	636.49	75.42	0.15
Math	45	632.31	65.00	45	628.04	66.13	0.07

Tables 14, 15, and 16, provide summary information for the three 2 x 2 x 5 mixed ANOVA's on the main and interaction effects for this study. Of interest in this study are the main effect on type of program and all interaction effects associated with type of program. As can be seen in Table 14, results indicated a significant main effect of type of program, $F(1, 324.73) = 20.71$, $p < .001$, for the reading area. They also indicated a non-significant interaction

effect of change over time x type of program (p = .960) and a non-significant three way interaction (p = .361) so the program was similarly effective over time. The main effect of time was further investigated by means of polynomial contrasts, indicating that there was a linear increase in CSAP Reading scores (p < .001). The quadratic and cubic trends were also significant at p = .049 and .025, respectively so the trend lines are not perfectly straight as can be seen in Figure 3.

The type of program main effect and the non-significant program type x primary language interaction (p = .831) indicate that students who attended the treatment

Table 14

Mixed ANOVA Results for CSAP Reading Achievement as a Function of Type of Program, Primary Language, and Change over Time

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,222,814.08	1.00	1,222,814,08	77.91	<.001
Error 1	1,396,961.89	89.00	15.66.20		
	Within subjects				
Type of Program	72,569.39	1.00	72,569.39	20.71	<.001
Type of Program x Primary Language	161.05	1.00	161.05	.05	.831
Change Over Time	266,002.21	3.01	88,287.87	104.80	<.001
Change Over Time x Primary Language	6,547.41	3.01	2,173.13	2.58	.054
Change Over Time x Type of Program	287.36	3.65	78.76	.14	.960
Change over Time x Type of Program x Primary Language	2,257.48	3.65	618.72	1.09	.361
Error 2	184,862.72	324.73	569.29		

program had higher CSAP reading scores across Junior High and High School when compared with their matched pairs, regardless of their native language. In other words, native English

speaking and native Spanish-speaking students benefit similarly and maintain their benefits from the treatment program across time (see Figure 3).

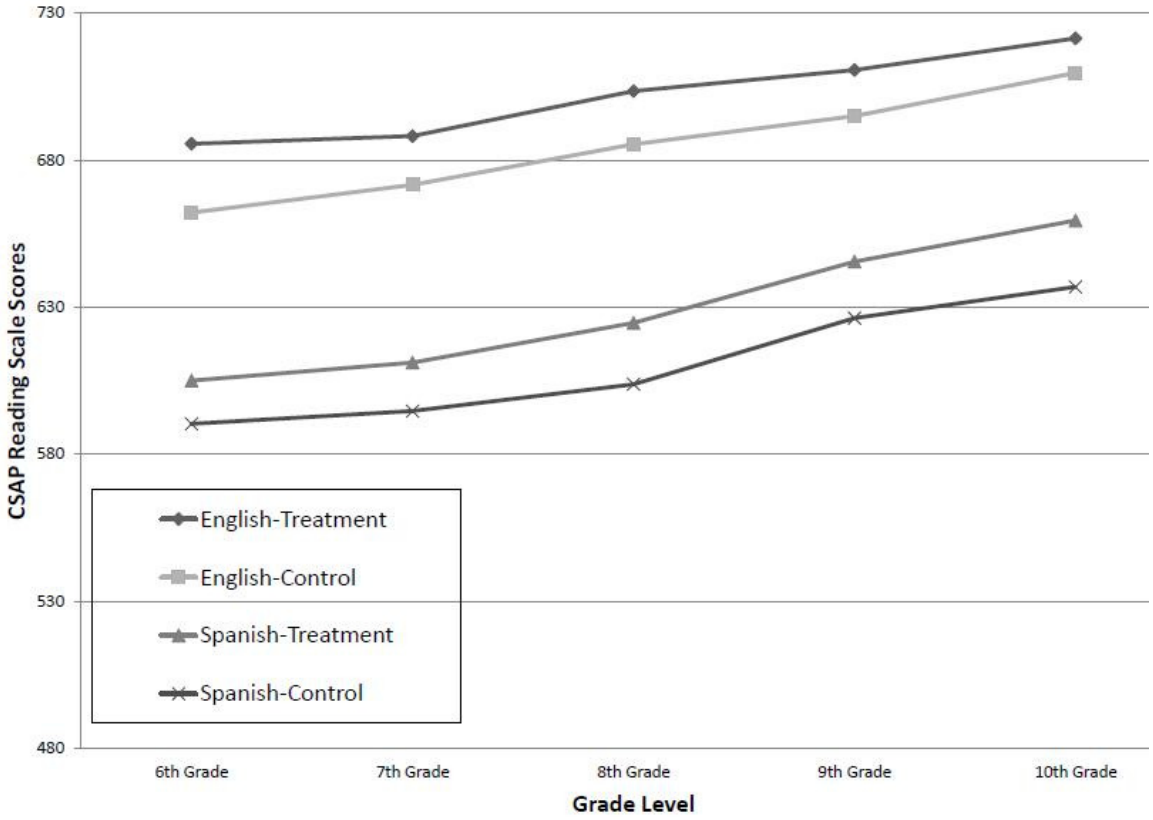

Figure 3. Comparison of average CSAP Reading scale scores by grade level, primary language, and treatment group.

The same pattern was observed for the writing area (see Table 15 and Figure 4), a significant main effect of type of program, $F(1, 283.47) = 4.83$, $p = .031$, is present for the writing subject. Again the interactions with type of program were not significant. The main effect of time was further investigated by means of polynomial contrasts, indicating that there was a linear increase in CSAP Writing scores ($p < .001$). The linear and cubic trends were also significant at $p = < .001$ and .014, respectively for the time x language interaction. These mean that the two languages have somewhat different straight and 2-bend lines as can be seen in Figure 4.

Table 15

Mixed ANOVA Results for CSAP Writing Achievement as a Function of Type of Program, Primary Language, and Change over Time

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,744,647.26	1.00	1,744,647.26	74.52	<.001
Error 1	1,896,277.42	81.00	23.410.83		
	Within subjects				
Type of Program	32,151.61	1.00	32,151.61	4.82	.031
Type of Program x Primary Language	7.16	1.00	7.16	.001	.974
Change Over Time	185,598.08	2.98	62,382.42	44.69	<.001
Change Over Time x Primary Language	38,451.88	2.98	12,924.28	9.26	<.001
Change Over Time x Type of Program	4,011.81	3.50	1,146.34	1.21	.308
Change over Time x Type of Program x Primary Language	3,556.16	3.50	1,016.14	1.07	.368
Error 2	269,229.51	283.47	949.75		

Results for math achievement followed the same pattern. Significant main effects of type of program, $F(1, 269.47) = 7.85$, $p = .006$ and relevant non-significant interactions are present for the math subject (see Table 16 and Figure 5). The main effect of time was further investigated by means of polynomial contrasts, indicating that there was a linear increase in CSAP Math scores ($p < .001$). The cubic trend was also significant at $p < .001$, so the trend lines are not perfectly straight as can be seen in Figure 5.

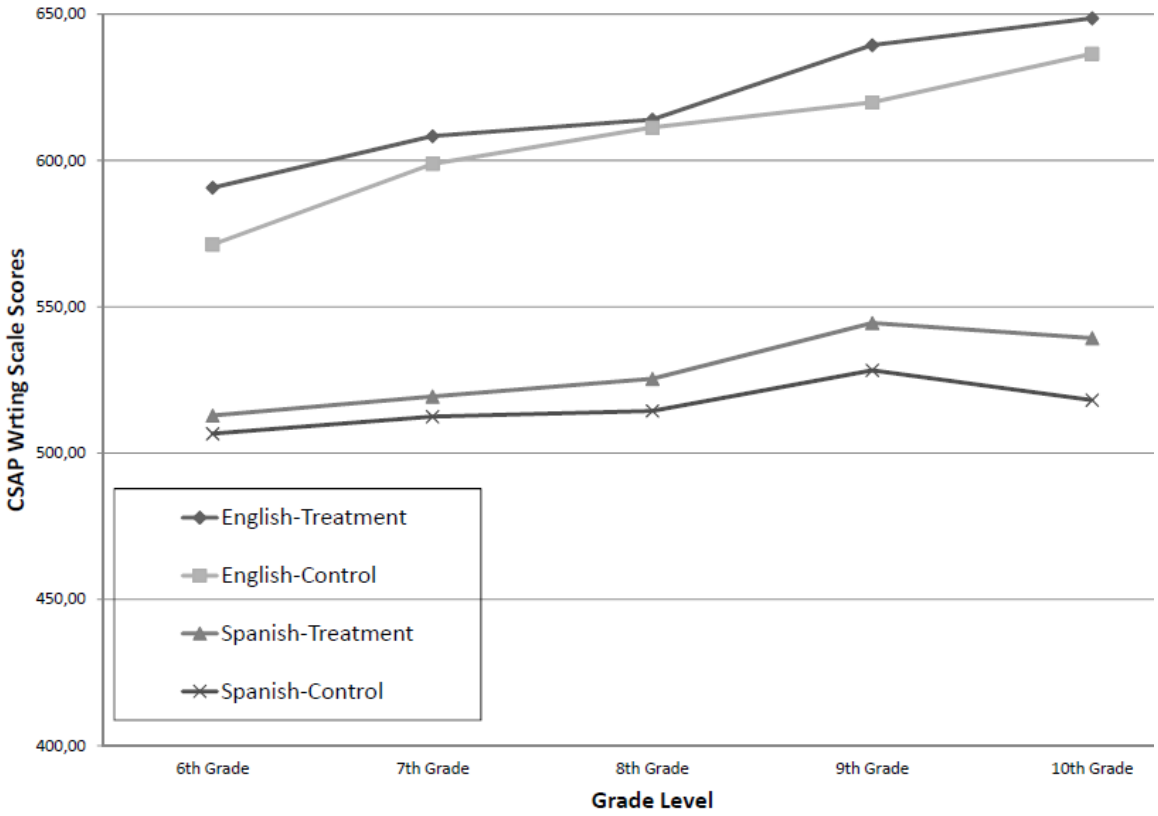

Figure 4. Comparison of average CSAP Writing scale scores by grade level, primary language, and treatment group.

With respect to our first research hypothesis, we observed a consistent main effect on program type across all three outcome domains, indicating the strength and breadth of the intervention across Junior High and High School.

Now we turn to the analysis of possible differential effects of the program type in each of the language groups across Junior High and High School.

Table 16

Mixed ANOVA Results for CSAP Math Achievement as a Function of Type of Program, Primary Language, and Change over Time

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,933,502.21	1.00	1,933,502.21	77.61	<.001
Error 1	2,167,522.48	87.00	24,914..05		
	Within subjects				
Type of Program	70,848.86	1.00	70,848.86	7.85	.006
Type of Program x Primary Language	29,166.37	1.00	29,166.37	3.23	.076
Change Over Time	188,860.31	3.58	52,711.81	56,86	<.001
Change Over Time x Primary Language	910.06	3.58	254.29	.27	.876
Change Over Time x Type of Program	1,207.12	3.10	389.73	.39	.765
Change over Time x Type of Program x Primary Language	4,070.09	3.10	1,314.07	1.32	.267
Error 2	267,529.38	269.47	992.81		

Hypothesis 2. CSAP Reading scores for native Spanish-speaking students in the treatment program and the control program were examined with a 2 x 5 (program type [treatment or control] x change over time [6^{th}, 7^{th}, 8^{th}, 9^{th}, 10^{th}]) mixed ANOVA. The ANOVA for reading revealed a significant type of program main effect, $F(1.00, 164.00) = 8.74$, $p = .005$ (see Table 17). There was no significant interaction effect ($p = .828$) between type of program and change over time. In addition to a highly significant linear increase over time ($p < .001$), the native Spanish-speaking group showed some evidence of a quadratic ($p = .05$) trend as can be seen in Figure 3. It is important that this analysis revealed an effect of type of program but no interaction with time. Thus, type of program effects on reading performance appeared similar over time for native Spanish-speaking students.

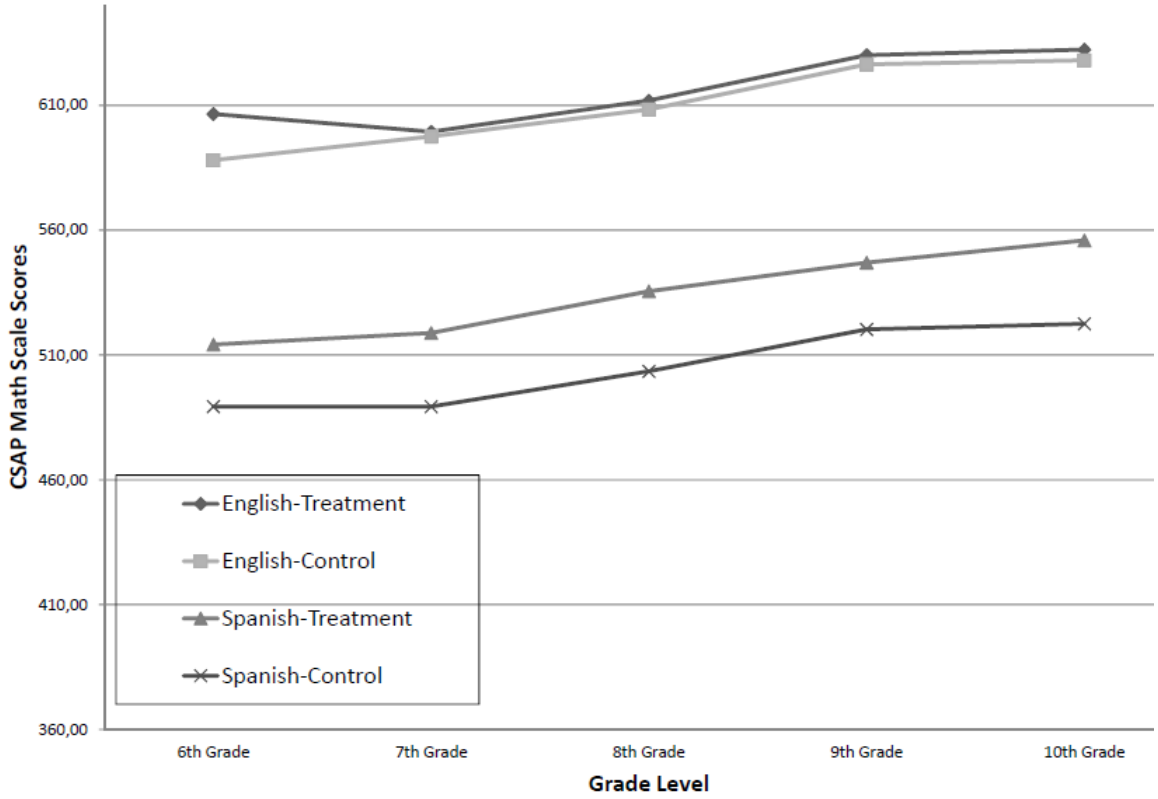

Figure 5. Comparison of average CSAP Math scale scores by grade level, primary language, and treatment group.

Table 17

Mixed ANOVA Results for CSAP Reading Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native Spanish-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	36,942.19	1.00	36,942.19	8.74	.005
Change Over Time	156,975.11	3.18	49,430.44	60.78	<.001
Change Over Time x Type of Program	856.99	4.00	214.25	.37	.828
Error	94,347.01	164.00	575.29		

CSAP Writing scores for native Spanish-speaking students in the treatment program and the control program were examined with another 2 x 5 mixed ANOVA. This ANOVA revealed a

non-significant type of program main effect, $F(1.00, 127.94) = 2.42$, $p = .128$ for the writing area (see Table 18). There was no significant interaction

Table 18

Mixed ANOVA Results for CSAP Writing Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native Spanish-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	15,055.29	1.00	15,055.29	2.42	.128
Change Over Time	35,499.64	2.67	13,319.94	10.31	<.001
Change Over Time x Type of Program	3,240.94	3.28	987.91	1.19	.316
Error	105,834.47	127.94	827.20		

effect ($p = .316$) between type of program and change over time. However, a significant linear increase over time ($p < .001$) was observed.

Even though CSAP average writing scores for native Spanish-speaking students were higher at each grade level than their matched counterparts, apparently, these differences in means were not large enough to reach significance, probably because of lack of power due to a relatively small sample of 40 in each group.

CSAP Math scores for native Spanish-speaking students in the treatment program and the control program were examined with another 2 x 5 mixed ANOVA. This ANOVA revealed a significant type of program main effect, $F(1.00, 125.83) = 8.69$, $p = .005$ for the math area (see Table 19). There was no significant interaction effect ($p = .813$) between type of program and change over time. In addition to a highly significant linear increase over time ($p < .001$), the native Spanish-speaking group showed some evidence of a cubic ($p = .044$) trend as can be seen in Figure 4.

It is important that the analysis revealed an effect of type of program but no interaction with time. Thus, type of program effects on math performance for Spanish-speaking students appeared similar over time.

Table 19

Mixed ANOVA Results for CSAP Math Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native Spanish-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	94,404.60	1.00	94,404.60	8.69	.005
Change Over Time	100,245.29	4.00	25,061.32	28.66	<.001
Change Over Time x Type of Program	1,115.52	2.93	381.21	.31	.813
Error	154,434.28	125.83	1,227.34		

According to these results, hypothesis two was confirmed for the subject areas of reading and mathematics, but not in the area of writing. Native Spanish-speaking students who graduated from the treatment program achieved significantly better in reading and math across Junior High and High School than the matched control group.

Hypothesis 3. The last part of the longitudinal data analysis in this study was done with the NES group. CSAP Reading scores for native English-speaking students in the treatment program and the control program were examined with a 2 x 5 (program type [treatment or control] x change over time [6^{th}, 7^{th}, 8^{th}, 9^{th}, 10^{th}]) mixed ANOVA. This ANOVA revealed a significant type of program main effect, $F(1.00, 192.00) = 12.36$, $p = .001$ for the reading area (see Table 20). There was no significant interaction effect ($p = .447$) between type of program and change over time. CSAP average reading scores for native English-speaking students were higher at each grade level than their matched counterparts (see Table 13 and Figure 3). In

addition, a highly significant linear increase over time ($p < .001$) was found for the native English-speaking group. All other polynomial contrasts with time were not significant.

It is important that the analysis revealed an effect of type of program but no interaction with time. Thus, type of program effects on reading performance appeared similar and stable over time for native English-speaking students.

Table 20

Mixed ANOVA Results for CSAP Reading Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native English-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	35,692.09	1.00	35,692.09	12.36	.001
Change Over Time	112,124.46	2.64	42,437.92	44.85	<.001
Change Over Time x Type of Program	1,757.09	4.00	439.27	.93	.447
Error	90,515.71	192.00	474.44		

CSAP Writing scores for native English-speaking students in the treatment program and the control program were examined with another 2 x 5 mixed ANOVA. This ANOVA revealed a non-significant type of program main effect, $F(1.00, 168.00) = 2.42$, $p = .127$ for the writing area (see Table 21). There was no significant interaction effect ($p = .348$) between type of program and change over time. Even though CSAP average writing scores for native English-speaking students were higher at each grade level than their matched counterparts (see Table 13 and Figure 4), these differences in means were not large enough to reach significance, a finding similar to that in the Spanish-speaking students.

Table 21

Mixed ANOVA Results for CSAP Writing Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native English-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	17,180.29	1.00	17,180.29	2.42	.127
Change Over Time	194,289.73	3.07	63,38.54	40.39	<.001
Change Over Time x Type of Program	4,367.76	4.00	1,091.94	1.12	.348
Error	163,395.04	168	972.59		

Remember, for both language groups combined (hypothesis 1), the program effect for Writing was significant at $p = .031$, but when the combined group was split, power was reduced.

CSAP Math scores for native English-speaking students in the treatment program and the control program were examined with another 2 x 5 mixed ANOVA. This ANOVA revealed a non-significant type of program main effect, $F(1.00, 121.61) = .64$, $p = .430$, for the math area (see Table 22). There was no significant interaction effect ($p = .189$) between type of program and change over time. CSAP average math scores for native English-speaking students were

Table 22

Mixed ANOVA Results for CSAP Math Achievement as a Function of Type of Program, Primary Language, and Change over Time for Native English-Speaking Students

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Within subjects				
Type of Program	4,601.60	1.00	4,601.60	.64	.430
Change Over Time	89,403.27	2.71	32,932.35	28.40	<.001
Change Over Time x Type of Program	4,196.30	2.76	1,518.23	1.63	.189
Error	113,095.10	121.61	929.96		

higher at each grade level than their matched counterparts (see Table 13 and Figure 5). It is important that the analysis revealed a lack of effect of type of program and no interaction with time.

According to these results, hypothesis three was confirmed. Native English-speaking students who graduated from the treatment program achieved as well as their matched counter parts in writing and math across Junior High and High School. Furthermore, in the reading area, native English-speaking students who graduated from the treatment program achieved significantly better than their matched group.

Figures 3, 4, and 5 clearly show that native English-speaking and native Spanish-speaking CSAP scores from the treatment program increase over time and are, on average, somewhat better throughout Junior High and High School in reading, writing, and math when compared with their matched samples. The same trend or pattern can be observed in Tables 12 and 13, which present means, standard deviations and effect sizes (ESs) associated with each of the simple effects regardless of statistical significance. Even though the overall main effect of program is small (Leech et al., 2005) in all three CSAP areas (reading, d = .28; writing, d = .16; and math, d = .22), a closer look allows for at least three interesting observations. First, ESs tend to be higher for native Spanish-speaking than for native English-speaking students in all three domains, and especially in grades 8, 9 and 10. (cf. ES for native Spanish-speaking students in Reading 9th grade, .52, with ES for native English-speaking students Reading 9th grade, .18). Second, ESs tend to get bigger for native Spanish-speaking students and smaller for native English-speaking students across Junior High and High School (time) in all three domains. Compare the ES for native Spanish-speakers 6th grade (d = .14) and 10th grade (d = .32) writing with the ES for native English-speakers 6th grade (d = .38) and 10th grade (d = .15) writing.

Compare also the ES for native Spanish-speakers 10th grade math (d = .50) with native English-speakers 10th grade math (d = .07). The same pattern can be observed for reading. Third, ESs for native Spanish-speaking students in math are the biggest ones at each grade level, with the only exception of 9th grade. This trend shows that the treatment program had its biggest effect in the math area for native Spanish-speaking students (see Tables 12 and 13). Overall, all three observations support hypotheses one, two, and three.

TWBI Students performance near the end of High School

With respect to our fourth research hypothesis, we expected to observe a consistent main effect on program type across all four ACT outcome domains, indicating the strength, breadth and durability of the intervention's effect. For hypothesis five we expected higher means for native Spanish-speaking students from the treatment group and a significant paired t test in each of the ACT four domains, English, reading, science, and math. With respect to our sixth research hypothesis, we expected to observe no significant differences in ACT means for each subject area (English, reading, science, and math) and across groups (treatment versus control). We expected none of the four paired t test to be statistically significant.

The ACT means, standard deviations, and sample sizes for the whole 11th grade group is presented in Table 23. The number of students included here (i.e., 64 matched pairs composed of

Table 23

Means and Standard Deviations for ACT 11th Grade Scores

	Treatment (n=64)		Control (n=64)	
Subject	Mean	Std. Dev.	Mean	Std. Dev.
English	21.33	7.05	18.91	7.37
Reading	22.53	6.87	20.19	6.66
Math	21.14	5.49	19.45	5.70
Science	21.22	6.16	20.28	6.19

17 native Spanish-speaking pairs and 47 native English-speaking pairs) is significantly lower than the number of students included in the previous analyses. This is due to the attrition that occurs in longitudinal studies. However, it should be noticed that the attrition rate is much higher for native Spanish-speaking than for native English-speaking students.

As can be observed in Table 23, all ACT mean scores are higher for students in the Treatment Program than in the Control Program. An examination for conformity to the assumptions underlying mixed ANOVA was conducted. Sphericity was not a concern because there were less than three levels in each of the factors.

Hypothesis 4. For the 11th grade ACT English scores, a 2 x 2 mixed ANOVA was computed. Both main effects were significant, however, only the type of program effect was of interest for this study. There was a statistically significant type of program effect, $F(1, 62) = 9.45$, $p = .003$ and a non-significant type of program by language effect, $F(1, 62) = .00$, $p. = .952$ (see Table 24).

Table 24

Mixed ANOVA Results for 11th ACT English Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,833.86	1.00	1,833.86	30.06	<.001
Error 1	3,781.88	62.00	61.00		
	Within subjects				
Type of Program	143.81	1.00	143.81	9.45	.003
Type of Program x Primary Language	.06	1.00	.06	.00	.952
Error 2	943.75	62.00	15.22		

Post-hoc comparisons of the means were not performed because there were fewer than three groups in each of the levels, and the direct observation of means allow to conclude that the performance of students from the treatment program was significantly higher.

The 2 x 2 mixed ANOVA of ACT reading scores yielded the same results, a statistically significant type of program effect, $F(1, 62) = 8.01$, $p = .006$; and a non-significant type of program by language effect, $F(1, 62) = .03$, $p = .967$ (see Table 25).

Table 25

Mixed ANOVA Results for 11th ACT Reading Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,433.94	1.00	1,433.94	27.07	<.001
Error 1	3,284.53	62.00	52.98		
	Within subjects				
Type of Program	135.31	1.00	135.31	8.01	.006
Type of Program x Primary Language	.03	1.00	.03	.00	.967
Error 2	1,047.19	62.00	16.89		

The third 2 x 2 mixed ANOVA of ACT math scores yielded also the same results, a statistically significant type of program effect, $F(1, 62) = 5.58$, $p = .021$ and a non-significant type of program by language effect, $F(1, 62) = .57$, $p = .454$ (see Table 26).

Table 26

Mixed ANOVA Results for 11th ACT Math Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	752.73	1.00	752.73	17.99	<.001
Error 1	2,593.99	62.00	41.84		
	Within subjects				
Type of Program	53.81	1.00	53.81	5.58	.021
Type of Program x Primary Language	5.47	1.00	5.47	.57	.454
Error 2	597.40	62.00	9.64		

The fourth 2 x 2 mixed ANOVA of ACT science scores yielded a different result, a non-statistically significant type of program effect, $F(1, 62) = .76$, $p = .388$ and a non-significant type of program by language effect, $F(1, 62) = .63$, $p = .432$ (see Table 27).

Table 27

Mixed ANOVA Results for 11th ACT Science Achievement as a Function of Type of Program and Primary Language

Source	Sum of Squares	*df*	Mean Square	*F*	*p*
	Between subjects				
Primary language	1,137.11	1.00	1,137.11	25.43	<.001
Error 1	2,771.89	62.00	44.71		
	Within subjects				
Type of Program	10.78	1.00	10.78	.76	.388
Type of Program x Primary Language	8.94	1.00	8.94	.63	.432
Error 2	883.94	62.00	14.26		

This result indicates that all students who attended the treatment program had significantly higher ACT English, reading, and math scores near the end of their high school when compared with their matched pairs. These results support hypothesis four in all ACT subject areas but science. Post-hoc comparisons of the means were not performed because there were fewer than three groups in each of the levels, and the direct observation of means allow to conclude that performance of students from the treatment program was better than their counterparts. Analysis by language group is presented in the next section.

Hypothesis 5. We performed four paired-samples *t* tests to determine whether the average ACT scores of native Spanish-speaking students from the treatment program were significantly different from that of the control program. The *t* tests confirmed a significant difference $t(16) = 2.68$, $p = .016$ only for the ACT Reading subject (see Table 28). This result only partially supported hypothesis five.

Table 28

Means and Standard Deviations for ACT 11th Grade Scores with t *Test Results for Treatment and Control Groups for Native Spanish-Speaking Students Only*

Subject	Treatment (n=17)		Control (n=17)		*t test*	*p*
	Mean	Std. Dev.	Mean	Std. Dev.		
English	15.00	3.48	12.65	4.77	1.80	.091
Reading	16.94	3.58	14.65	2.76	2.68	.016
Math	16.76	3.01	15.76	3.53	1.05	.311
Science	15.82	3.52	15.76	4.31	.05	.963

Remember, for both language groups combined (hypothesis 4), the program effect for English ($p = .003$), Reading ($p = .006$), and Math ($p = .021$) was significant at the 11th, but when attrition reduced the sample size substantially at 11th grade for the Spanish-speaking group, power was reduced.

Hypothesis 6. We also performed four paired-samples *t* tests to determine whether the average ACT scores of native English-speaking students from the treatment program were significantly different from that of the control program. The *t* tests confirmed a significant difference in three of the four ACT areas, [English, $t(46) = 3.02$, $p = .004$; Reading, $t(46) = 2.52$, $p = .015$; and Math, $t(46) = 2.93$, $p = .005$] (see Table 29). There was not a significant difference in the Science area. This result supported hypothesis six for all ACT areas: native English-speaking students from the treatment group performed equal or better than their matched counterparts.

Table 29

Means and Standard Deviations for ACT 11th Grade Scores with t *Test Results for Treatment and Control Groups for Native English-Speaking Students Only*

	Treatment (n = 47)		Control (n = 47)			
Subject	Mean	Std. Dev.	Mean	Std. Dev.	*t test*	*p*
English	23.62	6.61	21.17	6.84	3.02	.004
Reading	24.55	6.66	22.19	6.53	2.52	.015
Math	22.72	5.34	20.79	5.78	2.93	.005
Science	23.17	5.74	21.91	5.98	1.60	.117

Overall, students from the treatment program obtained mean ACT scores higher than the control group. The differences in English ($d = .34$), Reading ($d = .35$), and Math ($d = .30$) ACT results were not small but not medium either with the exception of Science ($d = .15$) where the difference was small (see Table 30).

Table 30

Effect Sizes for Main Effects and Simple Effects in English, Reading, Mathematics, and Science ACT Scores at 11th Grade Level

	English	Reading	Mathematics	Science
Main Effects Treatment vs Control	.34	.35	.30	.15
Treatment Spanish Speakers vs Control Spanish Speakers	.57	.72	.31	.02
Treatment English Speakers vs Control English Speakers	.28	.36	.35	.22

Effect sizes were medium and large for native Spanish-speaking students in English (d = .57) and Reading (d = .72) while they were small to medium for native English-speaking students in these areas (English, d = .28; Reading, d = .36), a pattern that is similar to the one that was observed in grades 6 to 10. Effect sizes in math were small to medium for both language groups, but small in science, especially for the native Spanish speakers.

Summary of Results

In support of our first research hypothesis, we found that students who attended the treatment program had significantly higher CSAP reading, writing and math scores at the end of their elementary school (6th grade) when compared with their matched pairs. We also observed a consistent main effect over time from 6th to 10th grade on program type across all three outcome domains, indicating the strength and breadth of the intervention across Junior High and High School. These results are noted in Table 31 along with those for hypotheses 2 and 3.

Hypothesis two was confirmed for the subject areas of reading and mathematics, but not in the area of writing. Native Spanish-speaking students who graduated from the treatment

program achieved significantly better in reading and math across Junior High and High School than the matched control group.

In regard to hypothesis three, Native English-speaking students who graduated from the treatment program achieved as well as their matched counter parts in writing and math across Junior High and High School. Furthermore, in the reading area, native English-speaking students who graduated from the treatment program achieved significantly better than their matched group.

We found that the overall program main effect is small in all three CSAP areas (reading, writing, and math), but a closer look allows for at least three interesting observations. First, ESs tend to be higher for native Spanish-speaking than for native English-speaking students in all three domains, and especially in grades 8, 9 and 10.

Table 31

Significance of Positive Program Effect for Each CSAP Subject in Grades 6 to 10

Subject	Whole Sample	Only NSS	Only NES
Reading	Significant (Table 14)	Significant (Table 17)	Significant (Table 20)
Writing	Significant (Table 15)	Not Significant (Table 18)	Not Significant (Table 21)
Math	Significant (Table 16)	Significant (Table 19)	Not Significant (Table 22)

Note. A Positive Program Effect refers to a mean score of students from the treatment program that is higher than the mean score of students from the control program. Significant or not significant refers to the result of the statistical test that was used in the study to test for differences in treatment conditions. NSS = native Spanish-speaking students; NES = native English-speaking students. The number after Table indicates the location of the table with the test that was used for each specific comparison.

Second, ESs tend to get bigger for native Spanish-speaking students and smaller for native English-speaking students across Junior High and High School (time) in all three domains. Third, ESs for native Spanish-speaking students in math are the biggest ones at each grade level,

with only the exception of 9th grade. This trend shows that the treatment program had its biggest effect in the math area for native Spanish-speaking students.

In regard to student´s achievement near the end of High School, results indicate that, on average, students who attended the treatment program performed somewhat better in ACT English, reading, and math scores when compared with their matched pairs. These results support hypothesis four in all ACT subject areas but science (see Table 31). Differences in English (d = .34), Reading (d = .35), and Math (d = .30) ACT results were not small but not medium either, but for Science (d = .15) where the difference was small.

Table 32

Significance of Positive Program Effect for Each ACT Subject

Subject	Whole Sample	Only NSS	Only NES
English	Significant (Table 24)	Not Significant (Table 28)	Significant (Table 29)
Reading	Significant (Table 25)	Significant (Table 28)	Significant (Table 29)
Math	Significant (Table 26)	Not Significant (Table 28)	Significant (Table 29)
Science	Not Significant (Table 27)	Not Significant (Table 28)	Not Significant (Table 29)

Note. A Positive Program Effect refers to a mean score of students from the treatment program that is higher than the mean score of students from the control program. Significant or not significant refers to the result of the statistical test that was used in the study to test for differences in treatment conditions. NSS = native Spanish-speaking students; NES = native English-speaking students. The number after Table indicates the location of the table with the test that was used for each specific comparison.

We found only partial support for hypothesis five. ACT Reading scores were significantly higher for native Spanish-speaking students than for their matched pairs (d = .72), but this was not the case for English, math and science. On the other hand, effect sizes were medium for native Spanish-speaking students in English (d = .57), and small to medium in Math (d = .31) suggesting a possible better performance of the treatment group. It seems that the lack of support for this hypothesis is related with the lack of power of the statistical test, given that

the number of native Spanish-speaking pairs was very small (n = 17). When using the whole sample, significant differences were found in all areas but science.

Our results supported hypothesis six for all ACT areas: native English-speaking students from the treatment group performed equal or better than their matched counterparts. Furthermore, students from the treatment program obtained mean ACT scores significantly higher than the control group in English (d = .28), reading (d = .36), and math (d = .35) but not science (d = .22).

It is interesting to compare the ESs for hypotheses 5 and 6 with those for 2 and 3. Effect sizes were medium and large for native Spanish-speaking students in English and Reading while they were small to medium for native English-speaking students in these areas, a pattern that is similar to the one that was observed in grades 6 to 10.

CHAPTER 5: DISCUSSION

Research findings in bilingual education are mostly unambiguous regarding the positive effects of bilingualism on children´s awareness of language and cognitive functioning (Bialysotck, 2001; Cummins, 2000). Nevertheless, bilingual education remains a highly debated issue in the U.S. and in other parts of the world. Controversy about the appropriateness and effectiveness of bilingual education has been a major focus of public debate (Bekerman, 2005).

This study explores the effects of a Two-Way Bilingual Immersion (TWBI) program on language majority and minority students. These effects need further explanation because studies of these types of programs are lacking. This is a methodologically sound study that examined achievement trends of language majority and minority students over eleven years with control for important student background characteristics.

The basis for conducting this study of a TWBI education program was: 1) to investigate the effects of the program on reading, writing, and math achievement in native English-speaking and native Spanish-speaking students, 2) to contribute to the base of knowledge on bilingual education in the U.S., and 3) to promote the use of methodologically sound comparative research in the bilingual education field.

The fundamental hypothesis was that the process of receiving instruction in two languages (English and Spanish) throughout elementary school (attendance at a TWBI school) would help the native Spanish-speaking students and not have a negative effect on the native English-speaking students in the performance of core academic areas (reading, mathematics, writing), and that this beneficial effect would carry through Junior High and High School in which instruction was delivered through a "business as usual" English-only model. This fundamental hypothesis shaped the following research hypotheses:

1. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in reading, mathematics, and writing at the end of 6th, 7th, 8th, 9th, and 10th grades than will a matched control group as measured by the Colorado Student Assessment Program test battery.
2. Those native Spanish-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in each of these achievement domains and each of these five years than a matched control group.
3. Those native English-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in each of these achievement domains and each of these five years as a matched control group.
4. Students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group as measured by the American College Testing test battery.
5. Those native Spanish-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve significantly better in English, reading, mathematics, and science in 11th grade than a matched control group.
6. Those native English-speaking students educated in a well implemented two-way bilingual immersion elementary school program will achieve at the same level in English, reading, mathematics, and science in 11th grade as a matched control group.

This is a longitudinal quasi-experimental study. The study design was an ex post facto, non-randomized, matched-pairs design. A multi-level matching procedure was used to match

students from the TWBI elementary school (treatment group) with comparable students from the school district (control group) beginning in third grade. Data on CSAP reading, writing, and math achievement were collected at sixth, seventh, eighth, ninth, and tenth grades. Data on ACT English, writing, mathematics, and science were collected at eleventh grade for each pair of matched students.

Data were collected through school district records on native Spanish-speaking and native English-speaking students from the two-way bilingual elementary school and their matched pairs who were students from comparable programs within the school district. Eleven annual cohorts of students from the treatment school were matched on a student-by-student basis on seven variables – cohort year, student's primary language, years of enrollment in the program, ethnicity, gender, socioeconomic status, and 3rd grade performance test– with comparable students from within the school district.

These eleven cohorts of 3rd graders were then tracked to the end of elementary school, middle and high school and measured on their reading, writing, and math achievement scores at each year. ACT scores were also collected in 11th grade. This *ex post facto* study involved an extraordinarily complex sampling process and necessitated a minimum of eleven years of cohort-level data collection in order to provide an adequate number of experimental and control students who stayed in their respective schools and in the same school district long enough and for whom complete outcome data were available in order to address the questions of interest in this study. Despite the inherent weaknesses in this *ex post facto* study and the small sample sizes for ACT data, we believe we can make some cautious statements about the effects of TWBI program on the short and longer term achievement of both native Spanish and English speakers.

Effectiveness of the Intervention Program through 10th Grade

In support of our first research hypothesis, we found that English and Spanish-speaking students combined, who attended the treatment program had significantly higher CSAP reading, writing, and math scores at the end of their elementary school (6th grade) when compared with their matched pairs. We also observed a consistent main effect across Junior High and through 10th grade on program type across all three outcome domains, indicating the strength and breadth of the intervention over time.

It appears that this well implemented TWBI program works and its effects apparently extend well beyond the intervention period into Junior High and High School. In recent years, quasi-experimental studies of TWBI programs have appeared in the literature with similar results. One outcome study of an enrichment bilingual education program that looked at achievement found that native English-speaking and native Spanish-speaking students did not become equally bilingual and biliterate, but they did outperform their peers in their first and second language by the upper elementary grades (Freeman, 1998). Studies by Castillo (2001), Coy and Litherland (2000), Lucido and McEachern (2000), Sera (2000), and Stipek, Ryan, and Alarcón (2001) focused on academic achievement of early elementary students who were enrolled in TWBI programs and consistently reported achievement levels for native English-speaking and native Spanish-speaking students in TWBI programs to be equal to or exceed achievement levels of their peers in elementary schools that offered other types of bilingual education programs.

Effectiveness of the Program Intervention by Language

Hypothesis two was confirmed for the subject areas of reading and mathematics, but not in the area of writing. Native Spanish-speaking students who graduated from or participated for

at least four years in the treatment program achieved significantly better in reading and math across Junior High and High School than the matched control group. A grade by grade, follow up analysis, comparing the performance of native Spanish-speakers with their matched peers, using paired *t* tests, showed statistically significant differences in reading and math in 8^{th}, 9^{th}, and 10^{th} grades, with students from the treatment program performing better than students from the control group. When comparing native Spanish speakers from the TWBI treatment program with native Spanish speakers who attended an ESL program (control group), the TWBI students outperformed the others in reading and math across Junior High and High School, and even in writing the TWBI Spanish-speakers had slightly higher mean scores at each grade level, with an effect size of d=.32 at 10^{th} grade.

In regard to hypothesis three, Native English-speaking students who graduated from the treatment program achieved as well as their matched counter parts in writing and math across Junior High and High School. Furthermore, in the reading area, native English-speaking students who graduated from or participated for at least four years in the treatment program achieved significantly better than their matched group. A grade by grade, follow up analysis, comparing the performance of native English-speakers with their matched peers, using paired *t* tests, showed statistically significant differences in reading in grades 6, 7, 8, and 9, with students from the treatment program performing better than students from the control group.

We found that the overall program main effect size (ES) is small in all three CSAP areas (reading, writing, and math), but a closer look allows for at least three interesting observations. First, ESs tend to be higher for native Spanish-speaking than for native English-speaking students in all three domains, and especially in grades 8, 9 and 10. Second, ESs tend to get bigger for native Spanish-speaking students and smaller for native English-speaking students across

Junior High and High School (time) in all three domains. Third, math ESs, for native Spanish-speaking students, are the biggest ones at each grade level, with only the exception of 9th grade. This trend shows that the treatment program had its biggest effect in the math area for native Spanish-speaking students.

This finding supports the assertion of Ramirez et al. (1991) that providing substantial instruction in Spanish speaking students' primary language does not impede their long-term achievement in any of the core academic areas. In fact, we found just the reverse.

Long Term Effectiveness of the Program Intervention

In regard to student´s achievement near the end of High School, results indicate that English and Spanish-speaking students combined who attended the treatment program performed significantly better in ACT English, reading, and math scores when compared with their matched pairs. These results support hypothesis four in all ACT subject areas but science. Differences in English (d = .34), Reading (d = .35), and Math (d = .30) ACT results were not small but not medium either; however for Science (d = .15) the difference was small and not significant.

We found only partial support for hypothesis five. ACT Reading scores were significantly higher for native Spanish-speaking students than for their matched pairs (d = .72). Although the means for English, math and science were somewhat higher for the treatment group, the difference was not significant. On the other hand, effect sizes were medium for native Spanish-speaking students in English (d = .57), and small to medium in Math (d = .31), suggesting a somewhat better performance of the treatment group. It seems that the lack of support for this hypothesis is related with the lack of power of the statistical test, given that the number of native Spanish-speaking pairs was very small (n = 17). When using the whole sample, both English and Spanish speakers, significant differences were found in all areas but science.

Our results supported hypothesis six for all ACT areas: native English-speaking students from the treatment group performed equal to or better than their matched counterparts. Furthermore, students from the treatment program obtained ACT mean scores significantly higher than the control group in English (d = .28), reading (d = .36), and math (d = .35); the means for science (d = .22) were not significant but in the same direction with the treatment group higher.

It is interesting to compare the ESs for hypotheses 5 and 6 with those for 2 and 3. Effect sizes were medium and large for native Spanish-speaking students in English and Reading while they were small to medium for native English-speaking students in these areas, a pattern that is similar to the one that was observed in grades 6 to 10.

Implications

The overall result of this study is aligned with Lindholm-Leary and Howard´s (2008) conclusions after examining the language and literacy development and math achievement of students in TWBI programs at secondary levels. In their review of the literature, they found that student achievement remains comparable through the secondary grades for both types (50/50 and 90/10) of TWBI programs. The authors concluded that native Spanish-speaking and native English-speaking students in TWBI programs perform at comparable or superior levels compared to same-language comparison peers:

> Findings are consistent across studies that included… students from different demographic backgrounds, and a variety of districts and states. In addition, the results are similar across longitudinal and cross-sectional data, with small-scale and large-scale studies and with research studies in various TWBI program environments (p.194).

The consistency of probability values and effect sizes for the main effect of the treatment program (TWBI) over the control ("business as usual" elementary school programs), as a whole,

and for both native English and Spanish speaking students suggests that TWBI programs, when implemented properly by schools, must be considered at least equally as effective in core academic achievement areas as "business as usual" elementary schooling, and is probably more effective in the long term.

This conclusion is supported by the pattern of findings in favor of dual language immersion programs and, more importantly, is completely independent of the beneficial effect of learning a second language for native English speakers and maintaining their first language for native Spanish speakers. Those particular beneficial effects were not investigated in this study. Given the sampling design for this study, however, this conclusion generalizes only to those students who stay with this schooling model throughout their elementary schooling, which is consistent with the research of Thomas & Collier (1997, 2001). Despite the limited sampling from only one TWBI school, this finding adds power to the theory that has built up with multiple small-scale studies of the effects of TWBI programs. Our findings about the TWBI program of this study resonate particularly well with the evaluation made by de Jong and Howard (2009):

> Studies have consistently shown that TWI students generally perform better than or equal to similar peers in non-TWI programmes on academic achievement measures…, though it has been noted that language minority students tend to perform below their fluent English peers within TWI programmes, even when controlling for students' free/reduced lunch status (p. 19)

Our study also lends support to Slavin and Cheung's (2005) meta-analytic review where they concluded that for Spanish speaking students, "rather than confusing children, as some have feared, reading instruction in a familiar language may serve as a bridge to success in English" (p. 274).

Very few studies have evaluated the long term effect of a TWBI program, and even fewer have used a longitudinal approach to evaluate the academic performance of native Spanish-

speaking and native English-speaking students in TWBI programs. Long term performance of native English-speakers in this study has shown to be similar to native English speakers from the TWBI Amigos program, instituted in Cambridge, Massachusetts, in 1986. After comparing native English speakers from the Amigos program with native English-speaking students from a mainstream program, Cazabon, Nicoladis, and Lambert (1998) found that "the English-Amigos are not behind in English, even though they receive only 50% of their instruction in English; their English seems to be as good as, or in many instances better than, that of students who are in an all-English program." (p.13)

Therefore, another implication of our results applies to parents and policymakers who have worried that placing native English speaking students in a TWBI program for the language benefit would detract from those students' long-term achievement in core academic subjects. Our findings, though generated by small sample sizes and an admittedly weak quasi-experimental design, suggest just the reverse. Native English-speaking students from the treatment group performed equal to or better than their matched counterparts in all ACT areas evaluated.

Although we did not test specifically in this study for reading and writing fluency in Spanish for these native English speaking students as they moved into Junior High and High School, anecdotal information acquired during this study indicates that these students had achieved high levels of such fluency, a genuinely important effect of the TWBI programs. The fact that this acquisition of both receptive and expressive language proficiency in Spanish can occur simply as a by-product of how instruction is delivered in elementary school is an important phenomenon that needs much more attention by researchers and policymakers in elementary education in the U.S.

It seems that the costs associated with implementing a TWBI differ from those of implementing a "business as usual" program in several areas. A significant amount of effort is involved in the implementation of a TWBI program. As local educators and policymakers ponder the significant investment in time, and personnel and financial resources associated with establishing a TWBI program, the concern for increased academic achievement and language fluency across all student populations must be part of the planning considerations. According to Howard and Christian (2002), many successful programs have found that some extra funding is necessary to provide staff development and purchase materials in the target language, especially for library and research materials. TWBI programs provide instruction in two languages to integrated groups of students, so it is a complicated and challenging model to implement effectively.

This longitudinal research on the effectiveness of a well implemented TWBI program has demonstrated that NSS students who attended the program tend to outperform those in the control group in their academic achievement in Junior High and High School. It seems that they also tend to complete their High School at a higher rate than their counterparts. This TWBI program not only promotes bilingualism by incorporating both the minority language and English into the academic setting, but has demonstrated its long-term beneficial effects on the academic achievement of its students. This form of Additive Bilingualism (Roberts, 1995) receives support from our findings, which could be used to support the theory of additive bilingualism.

Differential Attrition Rates

Our sample sizes were, at the older grades, very small. As shown in Table 4, a total of 82.5% of the native English-speaking students from the treatment program were included in the

study given that they met inclusion criteria (enrolled 4 or more years in the treatment program, not receiving special education services, and had demographic data available) versus only 69.2% of the Spanish-speaking students. This is known as treatment attrition (Shadish, Cook, & Campbell, 2008), and it had a different pattern by language group. It is probable that this phenomenon was the result of the characteristics of the Spanish-speaking population who attended the treatment program, comprised mainly of low-income immigrant families who tend to move more frequently.

As shown in Table 7, the number of students who completed the treatment program and were included in the study from the first five cohorts (96-97 to 01-02) was 230. Conversely, the number of students from the treatment program with valid achievement data in 11th grade was 109 (47%). This is known as measurement attrition (Shadish, Cook, & Campbell, 2008) because students completed the treatment program but there was a "failure" to complete outcome measures.

The analysis of measurement attrition allows us to make two observations. First, measurement attrition was higher for native Spanish-speakers in the treatment group. There were 61 out of 122 (50%) NES from the first five cohorts, who continued in the school district and were tested in 11th grade versus 48 out of 108 (44%) NSS. This pattern was also observed in the treatment attrition. Second, and more importantly, the measurement attrition in the control group was even more pronounced. As just stated, out of 108 native Spanish speakers from the treatment program, only 48 (44%) were in the school district at 11th grade. In contrast, from the 108 matched native Spanish speakers included in the study in third grade (control group), only 16 (15%) were in the school district at 11th grade (see Table 7).

We clearly had a very different pattern of measurement attrition for the native Spanish speakers in the control group. It appears that native Spanish speakers from the treatment program are more likely to stay in the school district than native Spanish speakers from other programs. This was an unexpected but important finding. It could be possible that native Spanish speakers who attended the treatment program received the benefits of a coherent and theory-based program that successfully helped them improve their academic achievement and allowed them to pursue and navigate their secondary level of instruction. Studying with native English-speaking students all day could be another factor influencing their performance.

The increased pattern of measurement attrition for the native Spanish speakers in the control group affected the power of statistical analysis because it reduced the number of participants included in the calculations. Overall, this pattern in the attrition does not affect the credibility of our findings, but make them more robust.

Limitations

The reader should be aware of several limitations to this study that may compromise the results and interpretations, and its generalizability to other school contexts. First are the inherent limitations to causal inferences that can be drawn from non-randomized, *ex post facto* studies like this one. This design represents one of the few ways whole school reform models can be studied, and the matched sampling design was implemented with as much attention to equating groups as was feasible. Nonetheless, cautious causal inferences are made in this study, and we recommend the reader interpret them as one piece of evidence in what needs to be a host of original research studies before any certainty about the effects of TWBI programs can be confirmed.

Second, there is a limitation to the interpretability of the study associated with the differential characteristics of the Spanish-speaking and English-speaking samples in the treatment groups. Almost all of the Spanish-speaking students came from low SES homes, while only about one fourth of the English speaking students came from such homes. This is simply an artifact of who attended the dual language immersion school in this study. Hence we could make no direct comparisons of English-speaking with Spanish-speaking students on achievement because of the SES confound and our inability to control for it through sampling, and no such comparisons across language groups should be inferred.

Another difference between NSS and NES who participated in the program is associated with the level of interest families show toward including their children in the TWBI program. Over the years, the school administration had to keep a waiting list for NES whose parents were very motivated to include their children in the program. On the contrary, the same administration had to make recruitment efforts to get enough NSS to register in the program. This is a limitation because this phenomenon was not occurring in the control group and there was no way to control for it. However, for NSS this phenomenon is not a limitation to the comparability of the two groups but makes our results more prominent given that NSS in the experimental group did not have the "special" motivation that NES families had, and still they outperformed their counterparts.

Third, and perhaps most importantly, the generalizability and perhaps the policy-related utility of our study is limited by our own explicit attention to treatment fidelity. We proactively limited our sample to only those children in both the dual language immersion school and their "business as usual" elementary school matched controls, who attended those schools for a minimum of four years and for most students their entire elementary school experience. We did this to confirm treatment effects under the assumption that sustained enrollment in the treatment

school was essential to ascertaining a legitimate treatment effect. However many students, particularly low SES Spanish-speaking students (which represented most of the native Spanish-speaking accessible population in this study), routinely migrate in and out of elementary schools and our study results cannot generalize to this substantial proportion of students.

An additional caution needs to be mentioned as readers review the results of this study. Because of large initial differences in socioeconomic status and achievement levels between native-Spanish speakers and native-English speakers, this study did not make comparisons across language groups; hence comparisons of native English-speaking students and native-Spanish speaking students, both of whom attended the treatment program, for example, were not made. In order to control for background characteristics that have been found to influence differences in achievement scores, in this study, native English-speaking students were compared only with native English-speaking students across programs and native Spanish-speaking students were compared only native Spanish-speaking students across programs.

Even thought data were collected over a period of more than eleven years, no evaluation of the fidelity of the treatment program implementation was carried out. Evaluating the characteristics of this 50/50 model, as well as teachers and staff members´ rotation across time could provide information to better explain the effects of the program.

Recommendations for Further Research

The quality of the design used in this study could be used to further the evaluation of TWBI programs. Replications of this study could be carried at other well implemented TWBI programs. A different approach could be taken and earlier cohorts from the treatment program could be compared with later cohorts, rather than the overall effect for all cohorts combined that has been shown so far.

This study relies wholly on school records, therefore cannot address the complex teacher and instructional factors that can influence student outcomes (Lindholm-Leary & Howard , 2008). Benefits of additive bilingualism were not assessed here, so including outcome measures to evaluate acquisition of a second language and bicultural attitudes could add an important perspective to the benefits already explored of TWBI programs.

REFERENCES

ACT (2012a). *ACT fact sheet*. Retrieved from ACT website: http://www.act.org/newsroom/factsheets/act.html

ACT (2012b). *The condition of college and career readiness 2012.* Retrieved from ACT website: http://media.act.org/documents/CCCR12-NationalReadinessRpt.pdf

ACT (2007). *The ACT technical manual*. Retrieved from ACT website: http://www.act.org/aap/pdf/ACT_Technical_Manual.pdf

Alanís, I. (2000). A Texas two-way bilingual program: Its effects on linguistic and academic achievement. *Bilingual Research Journal, 24*(3), 225-248.

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

August, D., & Hakuta, K. (Eds.). (1998). *Educating language-minority children*. Washington, DC: National Academy Press.

Baker, C. (*2006*). *Foundations of bilingual education and bilingualism*. (4th ed.). Clevedon, U.K: Multilingual Matters.

Baker, K.A. & de Kanter, A.A. (1981*). Effectiveness of bilingual education: A review of the literature.* Washington, D.C.: U.S. Department of Education, Office of Planning, Budget and Evaluation.

Bekerman, Z. (2005). Complex contexts and ideologies: Bilingual education in conflict-ridden areas. *Journal of Language, Identity, and Education*, *4*(1), 1–20

Black, M. (1991). Longitudinal studies in child maltreatment: Methodological considerations. In R. Starr (Ed.), (1991). *The effects of child abuse and neglect: Issues and research* (pp.129-143). London: Guilford

Bialystock, E. (2001). Bilingualism in development: Language, literacy and cognition. New York: Cambridge University Press.

Castillo, C.T. (2001). *The effects of a dual-language education program on student achievement and development of leadership abilities.* (Unpublished doctoral dissertation). Our Lady of the Lake University, San Antonio, TX.

Center for Applied Linguistics, (2011). *Directory of Two-Way Bilingual Immersion Programs in the U.S*. Retrieved from the Center for Applied Linguistics website: http://www.cal.org/twi/directory/

Cloud, N., Genesee, F., & Hamayan, E. (2000). *Dual Language Instruction*. Boston, Massachusetts: Heinle & Heinle.

Cobb, B., Vega, D., & Kronouge, C. (2009). Effects of an elementary dual language immersion school program on junior high school achievement. In: D.L. Hough (Ed), *Middle grades research: Exemplary studies linking theory to practice* (pp. 1-20). Charlotte, NC: Information Age Publishing

Cobb, B., Vega, D., & Kronouge, C. (2006). Effects of two-way dual immersion program on student performance: a longitudinal approach. *Middle Grades Research Journal. 1,* 27-47.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*.(2nd ed.). New York: Academic Press.

Collier, V. P., & Thomas, W. P. (2004). The astounding effectiveness of dual language education for all. *NABE Journal of Research and Practice, 2*(1), 1-20.

Colorado Department of Education. (2002). *CSAP technical information: Means, standards deviations and reliabilities.* Retrieved from Colorado Department of Education website: http://www.cde.state.co.us/cdeassess/as_TechInfo.htm.

Coy, S., & Litherland, L. (2000). *From a foreign language perspective: A snapshot view of a dual language program in two inner-city high poverty elementary schools.* (ERIC Document Reproduction Service No. ED 446450).

Crawford, J. (1989). *Bilingual education: History, politics, theory, and practice*. (4th ed.) Los Angeles: Bilingual Educational Services.

Cummins, J. (2000). Beyond adversarial discourse: Searching for common ground in the education of bilingual students. In P. McLaren, , & C.J.Ovando (Eds.), *The politics of multiculturalism and bilingual education: Students and teachers caught in the crossfire* (pp. 126-147). Madison, WI: McGraw-Hill, Higher Education.

Doherty, R.W., Hilberg, R.S., Pinal, A., & Tharp, R.G. (2003). Five standards and student achievement. *NABE Journal of Research and Practice 1*, 1-24.

De Jong, E. (2002). Effective bilingual education: From theory to academic achievement in a two-way bilingual program. *Bilingual Research Journal, 26*(1), 1-20.

De Jong, E.J., & Bearse, C.I. (2011). The same outcomes for all? High-school students reflect on their two-way Immersion program experiences. In D. J. Tedick, D. Christian & T.W. Fortune (Eds.), *Immersion education: Practices, policies, possibilities* (pp. 104-122). Clevedon, England: Multilingual Matters.

Edelsky, C. (1982). Writing in a bilingual program: The relation of L1 and L2 texts. *TESOL Quarterly, 16*, 211-228.

Enis, S. R., Ríos-Vargas, M., & Albert, N. G. (2011). *The Hispanic population: 2010* (2010 Census Briefs No. C2010BR-04). Retrieved from United States Census Bureau website: http://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf

Eisterhold-Carson, J., Carrell, P., Silberstein, S., Kroll, B., & Kuehn, P.A. (1990). Reading writing relationships in first and second language. *TESOL Quarterly, 24*, 245-266.

Escamilla, K. (2000). Teaching literacy in Spanish. In J. V. Tinajero, & R. A. DeVillar (Eds.), *The Power of two languages 2000: Effective dual-language use across the curriculum* (pp. 127-141). Farmington, New York: McMillan /McGraw-Hill School Division.

Freeman, R. (1998). Bilingual education and social change. Clevedon, England: Multilingual Matters.

García, E. E. & Jensen, B. (2010). Language development and early education of young hispanic children in the United States. In O. N. Saracho, & B. Spodek (Eds.), *Contemporary perspectives on language and cultural diversity in early childhood education* (pp. 43-64). Scottsdale, AZ: Information Age.

Genesee, F. (1987). Learning through two languages: Studies of immersion and bilingual education. Cambridge, MA: Newbury House.

Genesee, F. (Ed.). (1999). *Program alternatives for linguistically diverse students.* Santa Cruz, CA: Center for Research on Education, Diversity & Excellence.

Gilbert, S.M. (2001). *The impact of two-way dual-language programs on fourth-grade students: Academic skills in reading and math, language development, and self-concept development* (Unpublished doctoral dissertation). New Mexico State University, Las Cruces, NM.

Gomez, L., Freeman, D., & Freeman, Y. (2005). Dual language education: A promising 50-50 model. *Bilingual Research Journal, 29*(1), 145–164.

Greene J.P. (1997). A meta-analysis of the Rossell and Baker review of bilingual research. *Bilingual Research Journal, 21*(1) 1-18.

Hakuta, K., Bialystok, E., & Wiley, E. (2003). Critical evidence: A test of the critical period hypothesis for second language acquisition. *Psychological Science, 14*(1), 31-38.

Hobbs F., & Stoops, N. (2002). Demographic trends in the 20th century: Census 2000 special reports. Series CENSR-4. Washington, DC: U.S. Government Printing Office.

Howard, E. R., Sugarman, J., & Christian, D. (2003). *Trends in two-way immersion education: A review of the research* (Report No. 63). Baltimore, MD: Center for Research on the Education of Students Placed At Risk. Retrieved from: http://www.csos.jhu.edu/crespar/techReports/Report63.pdf.

Howard, E. R., & Christian, D. (2002). *Two-way immersion 101: Designing and implementing a two-way immersion education program at the elementary school level* (Educational Practice Report 9). Santa Cruz, CA and Washington, DC: Center for Research on Education, Diversity & Excellence. Retrieved from: http://www.cal.org/crede/pdfs/epr9.pdf.

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011). *Overview of race and Hispanic origin: 2010* (2010 Census Briefs No. C2010BR-02). Retrieved from United States Census Bureau website: http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Kortz, W. J., Jr. (2002). *Measuring the effects of the accelerated reader program on third grade English language learners' reading achievement in dual-language programs* (Unpublished doctoral dissertation). Sam Houston State University, Huntsville, TX.

Lanauze, M., & Snow, C. (1989). The relation between first and second language writing skills: Evidence from Puerto Rican elementary school children in bilingual programs. *Linguistics and Education, 1*, 323-339.

Leech, N., Barrett, K., & Morgan, G. (2005). SPSS for intermediate statistics: Use and interpretation. (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Leung, J. (2007). History and Curriculum Implications of Bilingual Education. Unpublished manuscript. Retrieved from http://javierleung.com/sub/docs/fall07/Bilingual%20Education.pdf

Lindholm-Leary, K. (2001). *Dual-language education*. Clevedon, England: Multilingual Matters.

Lindholm-Leary, K. J., & Howard, E. R. (2008). Language development and academic achievement in two-way immersion programs. In T. W. Fortune & D. J. Tedick (Eds.), *Pathways to Multilingualism: Evolving perspectives on immersion education* (pp. 177-200). Oxford, UK: Blackwell.

Love, S. (2005). Understanding mobile human-computer interaction. Burlington, MA: Butterworth-Heinemann.

Lucido, F., & McEachern, W. (2000). The influence of bilingualism on English reading scores. *Reading Improvement, 37*(2), 87-91.

McField, G. (2002). *Does program quality matter? A meta-analysis of select bilingual education studies*. (Unpublished doctoral dissertation). University of Southern California.

Menard, S. W. (2008). Introduction: Longitudinal research, design, and analysis. In S. Menard (Ed.), *Handbook of longitudinal research: design, measurement, and analysis* (pp. 3-12). Amsterdam: Elsevier.

Nieto, D. (2009). A brief history of the bilingual education in the Unites States. *Perspectives on Urban Education: 6*(1), 61-72.

Ovando, C. (2003). Bilingual education in the United States: Historical development and current issues. *Bilingual Research Journal, 27*, 1 – 24.

Paulston, C. B. (1992). *Linguistic and communicative competence: topics in ESL.* Clevedon: Multilingual Matters.

Ramirez, D., Yuen, S.D. & Ramey, D.R. (1991). *Final report: Longitudinal study of structured English immersion strategy, early-exit, and late exit transitional bilingual education program for language minority children, executive summary.* Washington, D.C.: Office of Bilingual Education.

Roberts, C.A. (1995). Bilingual education program models: A framework for understanding. *Bilingual Research Journal, 19,* 369 – 378.

Rolstad, K., Mahoney, K., and Glass, G. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy 19*(4): 572-594.

Rossell C.H. & Baker, K. (1996). The educational effectiveness of bilingual education. *Research in Teaching of English, 30,* 7 – 51.

Salazar, J.J. (1998). A longitudinal model for interpreting thirty years of bilingual education research. *Bilingual Research Journal, 22,* 1-12.

San Miguel, G. (2004). *Contested policy: The rise and fall of federal bilingual education in the United States, 1960-2001*. Denton, TX: University of North Texas Press.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Co.

Senesac, B. V. K. (2002). Two-way bilingual immersion: A portrait of quality schooling. *Bilingual Research Journal, 26*(1), 1-26.

Sera, G.L. (2000). *The nature and English language consequences of dual immersion schooling* (Unpublished doctoral dissertation). Indiana University, Bloomington.

Shneyderman, A., & Abella, R. (2009). The effects of the extended foreign language programs on Spanish-language proficiency and academic achievement in English. *Bilingual Research Journal, 32,* 241-259.

Slavin, R. & Cheung, A. (2005). A Synthesis of research on language of reading instruction for English language learners. *Review of Educational Research,75*(2), 247–284.

Soltero, S. W. (2004). *Dual language: teaching and learning in two languages.* Boston: Pearson/A and B.

Soltero, S. W. (2011). *Schoolwide approaches to educating Ells: Creating linguistically and culturally responsive K-12 schools*. Portsmouth, NH: Heinemann

Stipek, D., Ryan, R., & Alarcón, R. (2001). Bridging research and practice to develop a two-way bilingual program. *Early Childhood Research Quarterly, 16*(1), 133-149.

Swain, M., & Lapkin, S. (1991). Additive bilingualism and French immersion education: The roles of language and proficiency and literacy. In A. Reynolds (Ed.), *Bilingualism, multiculturalism, and second language learning: The McGill conference in honour of Wallace E. Lambert* (pp. 203-216). Hillsdale, NJ: Erlbaum.

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

Thomas, W.P. & Collier,V. (2001). *A national study of school effectiveness for language minority students' long-term academic achievement, final report, project 1.1.* Retrieved from http://www.crede.ucsc.edu/research/llaa/1.1_final.html.

U.S. Department of Education, Office of English Language Acquisition, Language Enhancement and Academic Achievement for Limited English Proficient Students, National Clearinghouse for English Language Acquisition. (2011). The growing numbers of English learner students, 1998/99-2008/09. Retrieved from: http://www.ncela.gwu.edu/files/uploads/9/growingLEP_0809.pdf

U.S. Department of Education. (2002). *Fact sheet: The no child left behind act.* Retrieved from: http://www.ed.gov/offices/OESE/esea/factsheet.html.

Willig, A. (1985). A meta-analysis of selected studies on the effectiveness of bilingual education. *Review of Educational Research, 55*(3), 269-318.