THE EFFECT OF INSTRUCTIONAL PROGRAM AND PHONOLOGICAL AWARENESS ON READING OUTCOMES AMONG EARLY ELEMENTARY SPANISH-SPEAKING ENGLISH LEARNERS

by
Karrie Aldrich Hilliard

A dissertation submitted to Psychological, Health and Learning Sciences,
College of Education
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in School Psychology

Chair of Committee: Dr. Jorge E. Gonzalez

Committee Member: Dr. Milena Keller-Margulis

Committee Member: Dr. Kristi L. Santi

Committee Member: Dr. Tammy D. Tolar

University of Houston
May 2020

Acknowledgement

This project was made possible through access to a deidentified dataset collected for the project entitled Oracy/Literacy Development in Spanish-Speaking Children, developed with funds from the National Institute of Child Health and Human Development (#HD-39-521). The principal investigator of this project was Dr. David J. Francis, Hugh Roy and Lillie Cranz Distinguished University Chair in Psychology at the University of Houston. I am grateful to Dr. Francis for the opportunity to add my study to the large and high-quality body of work this project has yielded over the last several years.

My gratitude extends further to my committee members, Dr. Jorge E. Gonzalez, Dr. Milena Keller-Margulis, Dr. Kristi L. Santi, and Dr. Tammy D. Tolar, who have shown support and encouragement throughout my journey. Dr. Tolar, in particular, has pushed me to be a more critical researcher from my first days as a graduate student. As my advisor and program director, respectively, Drs. Gonzalez and Keller-Margulis have provided excellent guidance in my development as a research-practitioner. Dr. Santi was instrumental in working through a large dataset of English Learners. I am thankful for the meaningful roles they played in this project and in my general doctoral training.

Abstract

Background: English Learners (ELs), the majority of whom are Spanish-speaking children, are at risk for academic underachievement in American public schools. Prior research on the effectiveness of instruction designed to support their learning, categorized as bilingual or immersion, is often confounded by selection bias and has not produced definitive conclusions. A recent study suggests that ELs with low English phonological awareness (PA) may benefit more from bilingual than immersion instruction in terms of reading achievement, but selection bias casts doubt on this relationship. **Purpose:** This study sought to examine if Spanish-speaking EL students with low English PA benefit more from bilingual than immersion instruction while controlling for selection bias using a matched-comparison group design. Preliminary analysis showed, however, that sample size-related challenges required inclusion of ELs of varying PA levels. It was hypothesized that PA would interact with instructional program, and that students with lower English PA would benefit more from bilingual than immersion. **Method:** From a sample of 689 Spanish-speaking ELs, matched-comparison groups (n = 45 per group) were formed using propensity score matching that improved balance between immersion and bilingual groups on beginning-of-first-grade English and Spanish reading, oral language, and PA. ANCOVA was used to evaluate if instructional program interacted with PA to influence end-of-second-grade reading achievement. **Results:** Interactions between instructional program and PA were not significant, nor were main effects, except for that of instructional program on Spanish reading, with bilingual students outperforming immersion students. A post-hoc power analysis indicated that the study was largely underpowered. **Conclusion:** Results suggest that bilingual education more

effectively promotes Spanish reading than immersion. Detection of effects in English reading was limited due to power, highlighting the need to take a more methodologically rigorous approach to studying this population, balancing priorities of detecting treatment effects and minimizing selection bias.

Table of Contents

List of Tables

Table Page

List of Figures

Figure Page

Chapter I:

Introduction

The educational achievement of the large and growing population of English Learners (ELs) in American public schools is gaining more attention from researchers and policymakers alike (Goldenberg, 2008; Rolstad, Mahoney, & Glass, 2005; U.S. Department of Education, 2010). Many of these students are Spanish-speaking, enrolled in bilingual programs, and struggling academically. Though most research supports bilingual instruction for ELs (August & Shanahan, 2006; Francis, Lesaux, & August, 2006; Thomas & Collier, 1997b), there is a long-standing lack of consensus in the field around instructional programming for ELs (National Academies of Sciences, Engineering, and Medicine, 2017; Rossell & Kuder, 2005).

Methodological challenges in studying the educational outcomes of ELs have slowed progress in understanding what instructional support benefits this predominantly Spanish-speaking yet highly heterogeneous population. English Learners enter American public schools with widely varying language ability profiles, and little is known about whether certain language ability profiles are better served by specific instructional programs (Cárdenas-Hagan, Carlson, & Pollard-Durodola, 2007). This study sought to examine whether there was a relationship between instructional program and reading achievement among early-elementary Spanish-speaking ELs with low English phonological awareness (PA), an early reading foundational skill, using a matched-case comparison design. However, the subset of students within the original sample with low English PA (n = 82) was relatively small and it was therefore not possible to find enough matched cases among the subset to produce a viable sample size for analysis. This

resulted in a decision to examine all students from the original sample, regardless of language ability profile. Despite the sample change, phonological awareness continued to be an important variable in the revised study, as it was examined for its possible interaction with instructional program in relation to reading outcomes.

Background

An English Learner (EL) is a student who is acquiring English as a second language. This term was adopted by many researchers in the 1990s because of its emphasis on the learner's growth and accomplishments rather than the deficit implied by such terms as limited-English-proficient (LaCelle-Peterson & Rivera, 1994). Most ELs are Hispanic from Spanish-speaking homes. During the 2015-16 school year, more than three-quarters of ELs were Spanish-speaking (Snyder, De Brey, & Dillow, 2019). Latinos hold a notable share of not only the EL population but also the general U.S. student population. Among school-aged children in the U.S., Latinos are the fastest-growing group: they represented 16% of the population in 2000 and 25% of the population in 2016 (National Center for Education Statistics, 2017).

Most Spanish-speaking ELs are served by schools in populous urban centers across the U.S., particularly in California, New York, and Texas, with nearly half of Latinos residing within 10 large metropolitan areas (Motel & Patten, 2012). Given that academic achievement in urban centers is often negatively impacted by inadequate resources and a lack of high-quality instructors (Halle, Hair, Wandner, McNamara, & Chien, 2012; Lonigan, Farver, Nakamoto, & Eppe, 2013), many Latino school-aged children and youth are negatively affected by these risk factors. When such geographic realities are combined with the additional challenge of learning a second language and

learning academic content, Spanish-speaking ELs face a unique challenge in their schooling, and often perform at a lower level than their English-speaking peers on important outcomes, including nationwide measures of academic achievement and graduation rates (National Center for Education Statistics, 2018; Ortiz, Valerio, & Lopez, 2012).

Only recently have EL-specific achievement data become accessible to researchers, driven primarily by increased standards of accountability for students across all sub-groups coming out of the 1997 and 2004 reauthorizations of the Individuals with Disabilities Education Act (IDEA), and the Adequate Yearly Progress (AYP) goals of the No Child Left Behind (NCLB) Act of 2001 (Goldenberg & Coleman, 2010; Waitoller, Artiles, & Cheney, 2010). Greater access to data and its analysis is promoting a better understanding of the instructional needs of ELs in public schools, but progress is slow and challenged by numerous factors. The heterogeneity of the population, the wide variability in instructional practices, the inadequacy of measurement tools, and socioeconomic confounds make the study of ELs' educational outcomes and the instructional programs that support them a formidable task. Briefly, instructional programs fall into two general categories: bilingual, which is characterized by native language instruction, and immersion, which is characterized by English-only instruction (Slavin, Madden, Calderón, Chamberlain, & Hennessy, 2011). On the question of language of instruction, inability to randomly assign students to different programs has made it difficult to identify a causal link between program and achievement (Nakamoto, Lindsey, & Manis, 2012). To further complicate matters, most researchers have concluded that the quality of instruction matters more than the language of instruction (L.

Q. Dixon et al., 2012; Gersten & Baker, 2000; Slavin et al., 2011). Furthermore, as is true for all emerging readers, it is important to screen ELs for early reading difficulties in order to intervene early and appropriately (Lonigan, 2006). Screening early may help in early identification of EL students in need of additional or deeper supports to forestall or prevent later reading challenges.

Recent years have seen an increased understanding of the utility of screening for reading difficulties (Johnson, Jenkins, Petscher, & Catts, 2009). Among the early literacy skills recommended for screening among ELs and their English-speaking peers is phonological awareness (PA; Leafstedt & Gerber, 2005; Manis & Lindsey, 2011). Defined as the ability to manipulate sounds by segmenting and blending them, PA is a prerequisite to reading (Anthony et al., 2011; Ball & Blachman, 1991; Gyovai, Cartledge, Kourea, Yurick, & Gibson, 2009; Lesaux & Siegel, 2003). Screening for difficulties with PA, including among ELs, can promote early intervention where needed and can decrease the risk of reading difficulties (Cirino et al., 2009). Because the process of learning to read by mapping sounds to symbols is fundamentally the same in both Spanish and English (Chiappe, Siegel, & Wade-Woolley, 2002; Ziegler & Goswami, 2005), and PA skills appear to transfer readily between the two languages (Branum-Martin, Tao, & Garnaat, 2015; Goodrich, Lonigan, & Farver, 2013), this screening approach has proved useful for early identification of reading difficulties and general reading readiness among Spanish-speaking ELs (Farver, Nakamoto, & Lonigan, 2007; Lesaux & Geva, 2006).

The purpose of this study was to evaluate whether ELs benefit from participation in a bilingual instructional program compared to an immersion instructional program in terms of reading growth in English and Spanish, and whether instructional program

interacts with phonological awareness. Gonzalez and colleagues (2015) found that Latino children from low-income households entered school at widely varying skill levels in terms of early reading and oral language skills, and therefore asserted that their instruction should be differentiated. More than a decade ago, Cardenas and colleagues (2007) observed that very few studies had examined the varying influence of language of instruction for different initial skill levels in students' native language and English, and the research gap still exists today. In a previous study examining the predictive validity of a Spanish/English PA screener in identifying students at risk of future reading difficulties in either language, results indicated that Spanish-speaking first-grade EL students with low PA in only English benefitted more from bilingual instruction over other instructional models in terms of reading outcomes in both languages by the end of second grade (Hilliard, 2016). These results suggested that the reading development of students with early signs of low English PA may be best supported in bilingual classrooms, and students with this ability profile could be identified through an early PA screening procedure. However, the methodological challenge of selection bias possibly explained the results, and so the current study sought to better control for the influence of selection bias and isolate the effect of the instructional program. A dictionary of defined terms relevant to this study can be found in Appendix A.

Present Study

Using an existing data set of over 2,000 Spanish-speaking ELs, 689 of whom met selection criteria, the present study set out to examine if students with the language ability profile of low English PA, as assessed through PA screeners in first grade, uniquely benefitted from bilingual instruction compared to immersion instruction in

terms of reading achievement in English and Spanish by the end of second grade. Implementing a matched-comparison group design to minimize the major methodological difficulty of selection bias, outcomes of students with low English PA were to be compared based on instructional program, further exploring the findings from the previous Hilliard (2016) study. However, the effects of selection bias were so significant, and the immersion and bilingual groups were so disparate in their initial language and literacy profiles, that a matching process with this subsample was not possible. The low English PA subsample included just 43 bilingual students and 39 immersion students, for a total of 82 students. Prior to the start of the intervention of immersion programming, these students self-selected into bilingual and immersion classrooms with widely varying reading, oral language, and phonological awareness skills in English and Spanish. The minimal overlap of the groups across multiple variables meant that even fewer students were available for viable matching, resulting in a final matched sample too small for analysis. As a result, the focus of the study shifted to consider the larger sample of 689 and whether an interaction between instructional programming and initial phonological awareness skills, or instructional programming alone, influenced their reading outcomes by the end of second grade.

The following research question and hypothesis were formed:

Research question. Do phonological awareness and instructional program interact to influence EL students' reading outcomes in English and Spanish?

Hypothesis. It was hypothesized that phonological awareness and instructional program would interact, and that the nature of that interaction would be that the lower the

phonological awareness, the greater the effects of bilingual instruction over immersion instruction on reading outcomes in English and Spanish.

Chapter II:

Review of the Literature

As the ethnic make-up of the United States continues to diversify, so do the challenges in educating American youth. Those students who enter American public schools with limited or no English skills, known and referred to here as English Learners (ELs), present a particularly complex and heterogeneous constellation of educational needs. The majority of ELs, 71%, come from Spanish-speaking Hispanic homes, most of which are of Mexican-descent (Ruiz Soto, Hooker, & Hatalova, 2015; United States Department of Commerce, 2012). Like their native-English-speaking peers, some Spanish-speaking ELs begin school with deficits in phonological awareness (PA; Lesaux & Geva, 2006). In English and Spanish, both alphabetic languages (Branum-Martin et al., 2015) with similar orthographies (Gottardo, Gu, Mueller, Baciu, & Pauchulo, 2011), PA provides a critical foundation for downstream reading skills like reading comprehension (Chiappe et al., 2002).

It is well-established that early intervention can alter the trajectory of children displaying signs of reading difficulty and prevent both the need for special education services (Foorman, Francis, Fletcher, Mehta, & Schatschneider, 1998) and the development of secondary behavioral and socioemotional problems (McBride & Siegel, 1997; Silver, 1989). The identification of both English-speaking and Spanish-speaking students at risk of reading difficulties using PA screeners is a method with growing support through research and in practice (Johnson et al., 2009; Majsterek & Ellenwood, 1995; Manis & Lindsey, 2011). The present study initially set out to examine whether a particular profile among Spanish-speaking EL students at risk of reading difficulties was

better served through bilingual instruction over English immersion instruction. The profile of interest was characterized by low English PA with average-or-above (i.e., not low) Spanish PA as assessed during first grade. The study shifted toward considering the way that PA may interact with instructional programming to affect reading outcomes. Despite the change, the underlying motivation of the study remains the same: knowing more about which students are best suited to benefit from one form of instruction over another could help inform decision-making around instructional programming for Spanish-speaking ELs and forestall or prevent the onset of entrenched reading difficulties that will ultimately require more intensive intervention to remediate.

This literature review will begin by reviewing the population growth and achievement levels of Spanish-speaking ELs, laws affecting these students, as well as the theories influencing their education in American public schools and the resultant instructional programming options offered to them. The review will then provide an overview of the relationship between phonological awareness and reading development, how oral language develops, and best practices in the assessment of Spanish-speaking ELs. The review will then transition to a focus on the importance of methodology in this study by providing some of the methodological challenges to conducting research with this population, encountered in the past by the author. A review of the literature around some popular methodological approaches chosen to address such challenges will follow. An overview of assessing causal inference, first through randomized controlled trials and then in the case of observational studies, will lead to a discussion of various methods for controlling group differences, such as propensity score matching. With the establishment

of the utility of propensity score matching as a method in observational studies, the current study will be introduced.

A Growing Population

Given the rate at which the number of ELs in American public schools is rising, delivering empirically-supported instruction for best possible educational outcomes for ELs is critical and urgently needed. During the 2015-16 school year, 9.5% of public school students were participating in programs for ELs, up from 8.1% in 2000 (Snyder et al., 2019). By 2050, it is projected that 34% of American children under the age of 17 will be immigrants or will have at least one parent who is an immigrant (Passel & D'Vera, 2008), increasing the demand on public schools to serve students from culturally and linguistically diverse backgrounds.

Most of these ELs are and will continue to be from Spanish-speaking homes. During the 2015-16 school year, 48.9% of children enrolled in American public schools were white and 25.9% were Hispanic, and by 2027, the proportion of Hispanic students is projected to rise to 29.2% (Snyder et al., 2019). More specifically, as many as 70% of Hispanic students are of Mexican descent (United States Department of Commerce, 2012). Based on these population estimates, the families of most ELs are native to Mexico and Spanish-speaking; however, even among Spanish-speaking ELs, there is a high degree of heterogeneity and consequent challenges in understanding and meeting their educational needs.

The Achievement Gap

Relative to their English-speaking peers, ELs face greater risk for low educational achievement. Risks include reading difficulties, the repeating of grades, and dropping out

of school (Artiles, Rueda, Salazar, & Higareda, 2005; August & Hakuta, 1997). Identification under the federal category of "limited English proficient" (LEP) is associated with the risk of dropping out of high school. Compounding high-risk factors include the high likelihood that the student will come from a low socioeconomic status (SES) home and an immigrant family (Callahan, 2013). The National Assessment of Educational Progress (2019) report notes that Hispanic fourth-grade students averaged a score of 209 on reading compared to 230 for white students, a 21-point difference. Among those students classified as EL, the average reading score was 191, whereas the average score of non-EL students was 224. For context, the cut score for fourth-grade reading students to achieve "NAEP Proficient" status (i.e., "representing solid academic performance") was 238. The low achievement trajectory can be traced back to children's first school experiences; both LEP status and low SES have been linked to a risk of reading difficulties as early as pre-kindergarten (Lonigan et al., 2013).

Socioeconomic status (SES) is a particularly important risk factor to consider when evaluating the academic achievement of ELs (Halle et al., 2012), though SES is difficult to disentangle from other risk factors, such as non-dominant language status (Lonigan et al., 2013). According to U.S. Census Bureau data (2018), as of 2017, a quarter of Hispanic children in the U.S. were living in poverty. This suggests that many Hispanic students, thus many ELs, come from poor families, given that 77.1% of students participating in EL programs speak Spanish at home (Snyder et al., 2019). Research has shown that children growing up in lower-SES homes experience fewer opportunities for language and print exposure and delayed vocabulary development when compared to their higher-SES peers (Hart & Risley, 1995; Hoff, 2003). A more developed Spanish

vocabulary can promote educational achievement in English (Proctor, August, Carlo, & Snow, 2006), and an underdeveloped home language vocabulary may place ELs at a disadvantage. This, combined with the fact that EL children are less likely to enroll in preschool than their English-speaking peers (Capps et al., 2005), suggests that EL children are at greater risk of failure in school before it has begun.

In summary, Spanish-speaking ELs face many risk factors that put them in danger of school failure. The rapid growth of ELs, Spanish-speakers in particular, highlights the need for focused efforts and resources aimed at mitigating the documented detrimental effects of low-SES and second language learning, yet low achievement outcomes indicate that many ELs do not receive the support they need.

Laws and Policies that Affect ELs

The achievement gap that persists today has a long history in ELs' educational experience, and past legislative efforts at the state and federal level have attempted to address it. The first major legislation to remedy educational disparities for ELs was the 1968 Bilingual Education Act (P.L. 90-247), though it did not require schools to use native language instruction. Modestly funded, the bill was intended to help poor Mexican-American children to learn English, and was named after the children it endeavored to serve, not the instructional program they were to experience (Wiese & García, 1998). Further increasing protections for ELs though again stopping short of a prescriptive instructional mandate, the landmark decision of *Lau v. Nichols* (1974) established language minority status as a viable basis for claims of discrimination in public schools and other institutions in receipt of federal funding. With the Supreme Court's declaration that ELs whose schools make no provisions for their language-related

needs are “effectively foreclosed from any meaningful education” (1974), ELs were henceforth entitled to differential treatment based on their language background (Wiese & García, 1998).

The reauthorizations of the Bilingual Education Act through the mid-1970s and 1980s gradually shifted support away from native language instruction, which saw a short-lived boon following the Lau v. Nichols decision, and provided more funding to English-only programs (Fitzgerald, 1993). Fitzgerald further argued that a rising population of Spanish-speaking immigrants and a wariness of U.S. involvement in world affairs led to a rise in nativism among Americans during this era. As a result, English-only lobbyists, such as U.S. English, and neo-conservative thinkers who viewed bilingualism as damaging to American society, began to gain traction (Cummins, 2001). As an example, Schlesinger (1991, p. 109) warned that “institutionalized bilingualism remains another source of the fragmentation of America, another threat to the dream of ‘one people.’” Though key provisions have since been overturned (Ulloa, 2016), it was this argument that galvanized support around the near-elimination of bilingual education with the passage of California Proposition 227 in 1998 (Olson, 2007) and fostered similar legislation in several other states.

California’s Proposition 227 imposed on local education agencies the onerous criteria that, in order for bilingual education to be offered, parents of at least 20 students were required to visit the school and waive rights to an English-only education annually, and the schools were required to also offer mainstream English and structured English immersion instruction (Unz & Ruchman, 1997). Following the implementation of the proposition, many schools were unable to comply, and the proportion of ELs receiving

bilingual instruction dropped from 30% to 8% in less than five years (Parrish et al., 2006). In 2016, California voters passed the Multilingual Education Act, eliminating the requirement for parent-signed waivers and returning more programmatic control to schools (Ulloa, 2016).

In contrast, in Texas, the offering of bilingual education to elementary-aged ELs has been required by law since 1981 per Senate Bill 477 (Truan & Vale, 1981). The bill also established guidelines for student identification for, placement in, and exit from services related to LEP status. Though recent years have brought about some policy changes to address lower academic achievement among ELs in Texas, the provision of bilingual programming, along with immersion programming options, has remained constant (Cortez & Johnson, 2008). These examples of California and Texas are relevant to the present study but also representative of types of state-led legislative shifts that directly impact educational options for ELs.

Theories Driving EL Instruction

Federal and state legislative efforts suggest that decisions about how to educate ELs in American public schools have been subject to sociopolitical movements and public opinion. While some view bilingualism as virtuous, others see it as damaging to American society (Cummins, 2001). Thus, instructional programming is often driven by district resources and political will (Fitzgerald, 1993) rather than evidence-based practices (Slavin et al., 2011). The proceeding paragraphs outline several prominent theories that attempt to explain the way that ELs develop a second language and the type of instruction that would most benefit them: cognitive academic language proficiency, the linguistic interdependence theory, and the bilingual advantage.

First introduced almost 40 years ago, the concept of cognitive academic language proficiency (CALP; Cummins, 1979b) is important because it draws the distinction between academic fluency and conversational fluency (basic interpersonal communicative skills, BICS). Understanding that these are two distinct skill continua assist in limiting the over-estimation of an EL's academic abilities when the child is able to converse in the second language in familiar, casual situations. The developmental timeline for each is quite different; research shows that second language oral proficiency (BICS) may take 3-5 years to develop, whereas academic proficiency (CALP) may take 4-7 years to develop (see Hakuta, Butler, & Witt, 2000). More recently, Cummins (2001, 2013) introduced a third component of language proficiency called discrete language skills, which are the set of rules governing language, such as phonology and grammar, and are acquired through instruction and practice, developing concurrently alongside BICS and CALP. Some discrete language skills are acquired early in schooling and some develop later. These three language domains are important to measure separately to gain a true understanding of an EL student's burgeoning abilities.

The linguistic interdependence theory (Cummins, 1979b) suggests that a common underlying proficiency explains the development of CALP in a second language. This is based on the assumption that cognitive/academic knowledge is transferable between similarly structured languages (like English and Spanish), and may help explain why older children initiating second-language-learning develop CALP in their second language (L2) more quickly than younger children, who have had less time to develop CALP skills in their first language (L1) (L. Q. Dixon et al., 2012). The linguistic interdependence theory supports the premise of bilingual education, valuing time spent

on developing a strong L1 foundation and challenging the tendency to devalue a minority language and culture (Cummins, 2009).

A third important theory is that the acquisition of a second language bestows upon the learner a set of cognitive and linguistic benefits beyond the inherent value of the ability to communicate in more than one language (Paradis, Genesee, & Crago, 2011a). Bilinguals have demonstrated greater abilities in selective attention and inhibitory control (Bialystok & Shapero, 2005), and these executive control skills are strongest when the bilingual individual has balanced proficiency between L1 and L2 (Yow & Li, 2015). Increased metalinguistic abilities may be explained by the structural sensitivity theory, which posits that the experience of knowing two languages improves the development of phonological awareness because the learner must break down and use phonetic units in a more variable context (Kuo, Uchikoshi, Kim, & Yang, 2016). Phonological awareness (PA) is the ability to manipulate sounds by segmenting and blending them, and is a prerequisite to reading (Anthony et al., 2011; Ball & Blachman, 1991; Gyovai et al., 2009; Lesaux & Siegel, 2003).

Current Educational Landscape for ELs

Primarily because program offerings and placement decisions may be more influenced by politics and available resources than by sound empirical evidence (Goldenberg & Coleman, 2010), Spanish-speaking ELs attending American public schools are served on a continuum of instructional settings. Such variety is also a result of a long period of equivocal research on outcomes of bilingual education (see Abedi, 2004); however, the efficacy of common instructional programs serving Spanish-speaking children has come under rigorous review in the last 15 years, and more is

known about the potential for long-term educational outcomes for students in each of these instructional settings.

Options may be grouped under the general headings of "bilingual" or "immersion" classrooms, designed to develop both the native language (L1) and English (L2) in the former, and only English in the latter. There are three common models of bilingual education: early transition bilingual, late transition bilingual, and dual language. There are two categories of common models of immersion education: traditional immersion and pull-out immersion (Ochoa, 2005). Each of these instructional programs is practiced in American public schools in some form and shares the aim of promoting growth in English language skills, either in the short or long term. Therefore, the challenge of acquiring English becomes a notable variable in the educational experience of all ELs. The question of how best to support students faced with this added challenge of acquiring English while learning the content of instruction has drawn scrutiny, and each program design has its own proponents and critics. A brief description and cursory overview of the most relevant research is provided below.

Early transition and late transition bilingual. Bilingual programs differ from other instructional programs in that a significant portion of the instruction is provided in L1 before transitioning to a greater focus on English, which may happen earlier or later in the educational career of the student (see Slavin et al., 2011). Early transition, also known as "early-exit" or "transitional," is typically offered for the first two-to-four years of school to a classroom of EL students. The philosophy driving this model is that children from culturally and linguistically diverse backgrounds need some time to adjust, but the goal must be to reach English proficiency as quickly as possible. Therefore, Spanish

reading instruction is provided early on as a bridge to English reading, but Spanish may not appear in other content areas outside of reading.

Late transition, also known as "maintenance," "late-exit," or "developmental," is typically offered for the first four-to-six years of school to a classroom of EL students. The model's philosophy is that a solid foundation must first be established in the child's native language such that academic skills may be readily transferable to English once second-language skills develop (Genesee, Paradis, & Crago, 2004). Another fundamental belief underlying the model is that there are cognitive benefits (Bialystok, 2010; Christoffels, de Haan, Steenbergen, van den Wildenberg, & Colzato, 2015) and economic value (Saiz & Zoido, 2005) in training students to be true bilinguals. Instruction in the early years is delivered in Spanish with some English language instruction and, over time, the balance gradually shifts to primarily English.

Transitional bilingual programs, whether early or late, gained more support after a national synthesis panel shared that proficiency in native language strongly predicts English language development (August & Hakuta, 1997). While acknowledging the complexity of the research, the members of the National Literacy Panel on Language-Minority Children and Youth concluded that bilingual education is preferable to an English-only approach for ELs because a preponderance of evidence suggests it has small to moderately-sized effects promoting English reading outcomes (August & Shanahan, 2006). A recent summary of extant research on bilingual education asserted that it produced superior reading outcomes in English compared to immersion settings (Goldenberg, 2013). The hypothesis of cross-linguistic transfer aims to explain these

findings (Durgunoğlu, Nagy, & Hancin-Bhatt, 1993; L. M. López & Greenfield, 2004), and is discussed further below.

Dual language. Dual language is the third variation on bilingual programs and is also referred to as "two-way" programming. Equal numbers of both English-speaking and EL students endeavor to become bilingual over the course of at least the first four-to-six years of schooling in this model. Like late transition, this model places value on bilingualism but adds a bicultural element with the presence of children from the dominant culture. Among a variety of dual language models is that in which the language of instruction varies by content area between the ELs' home language and English, achieving a 50-50 balance (Goldenberg & Coleman, 2010). Despite great variation in implementation, the common element is that reading instruction is provided in both languages for all students (Genesee et al., 2004).

This model is grounded on the theory that second language acquisition is a lengthy process (see Hakuta et al., 2000), and that there is a common underlying proficiency that can promote the transfer of skills between languages (Cummins, 1979b). Because it does not segregate ELs from mainstream students, it is more inclusive and strives to balance the social, academic, and language development of its participants (Christian, 1996). A highly cited longitudinal study of EL achievement across instructional settings was conducted by Thomas and Collier (1997a) and found that dual language programs yielded the highest English-reading scores by high school, relative to other instructional programs.

Traditional immersion and pull-out. Traditional immersion, also known as "content-based" or "sheltered English," focuses on delivering academic content in

English but an accessible manner, using gestures and other visual cues to assist. It is typically offered as needed, regardless of age. Models vary from students spending half to entire days in this setting, and the goal is English acquisition; native language maintenance is not facilitated. This approach is based on an assimilationist perspective, which promotes a maximum exposure theory, arguing that children deficient in English must receive as much English instruction as possible, as early as possible (Wiese & García, 1998).The difference between pull-out immersion and traditional immersion is that the instruction delivered in pull-out is explicit English language instruction, not content-based. This instruction occurs outside of the classroom and is supplemental to the child's regular instruction. This program may be available only to "newcomers" to the country, regardless of age, to ease their transition to American schooling, and typically excludes American children from Spanish-speaking homes.

In summary, school districts serve EL students along a continuum of instructional placements that, at one end, place greater emphasis on supporting L1 (bilingual) and, at the other, place a greater emphasis on promoting L2 (immersion). Notwithstanding varied views on model benefits, a recent review of evidence by Cheung and Slavin (2012) suggests that it is the quality of instruction that will most influence student outcomes. The review indicated that a variety of evidence-based approaches, in either bilingual or immersion settings, may work equally well in achieving the goal of English reading proficiency.

Pedagogical Challenges. No matter the instructional model, providing quality education to ELs is challenging, not only due to the perennial shortage of highly trained teachers of ELs, but also because of the complexity of the work faced by teachers once

they are hired. As explained by Cummins (2001), EL instructors are often plagued by two misperceptions about the nature of language proficiency. The first is that English ability is indicative of a child's general intellectual ability. The second is that a child's conversational skills in English provide a valid index of his overall mastery of the language. In addition to a risk of misunderstanding students' language abilities, teachers of ELs often appear to struggle to balance the pedagogical priorities of development of language and development of content knowledge (Gersten & Baker, 2000). One way to target the development of language is through explicit PA instruction, which can be very effective in shoring up PA deficits. In their meta-analysis of extant reading research, the National Reading Panel (2000) found that direct, early training in PA yielded a large effect size ($d = 2.37$) in PA development. Lesaux (2013) argues that ELs require explicit instruction around early skills-based competencies, such as those within phonological awareness.

There is evidence that PA skills are being explicitly taught, aligned with the phonics-based tradition, and that doing so yields great benefits to ELs and native speakers alike (August & Shanahan, 2006; Gersten & Baker, 2000; Lesaux, 2013). Such instruction was listed as one of the five essential pillars of reading by the National Literacy Panel on Language-Minority Children and Youth in 2006 (August & Shanahan). Research by Goldenberg and colleagues (2014) indicates that explicit instruction at the phoneme level is particularly helpful to students learning to read in English because of the complex orthography of the language. These students should receive explicit instruction in PA and regular assessment of PA skills to monitor progress (Lesaux, 2013).

Across instructional settings for ELs, it appears that teachers may neglect to provide explicit instruction promoting English oral language skills (Chiappe et al., 2002). Doing so is most important when there is not an immersive L2 environment found in the community (L. Q. Dixon et al., 2012), which is frequently true for Spanish-speaking ELs in the U.S. It is helpful to students to develop English oral language skills in the context of literacy instruction (Snow, 2006) but oral language skills should be targeted directly (Cárdenas-Hagan et al., 2007), with no assumption that sheltered instruction (where English oral language skills develop while learning science or social studies content) is sufficient (Gersten & Baker, 2000). Research shows that students require targeted instruction in vocabulary, listening comprehension, and reading comprehension (Manis & Lindsey, 2011), as well as structured opportunities to practice speaking English (L. Q. Dixon et al., 2012; Gersten & Baker, 2000; Graves, Gersten, & Haager, 2004).

Phonological Awareness: A Foundation for Reading

Phonological awareness (PA) is the knowledge that words are composed of sounds and has been defined as "the ability to consciously attend to the sounds of language as distinct from its meaning" (Lesaux & Geva, 2006, p. 55). Related skills include the ability to manipulate individual sounds of words by adding them together, deleting a sound, or re-ordering sounds (Paradis, Genesee, & Crago, 2011b). A precursor to reading (National Institute of Child Health and Human Development, 2000), PA is particularly relevant to the discussion of literacy development in Spanish-speaking ELs because of the transferability of PA skills from Spanish to English, as further discussed below.

In general, PA skills are essential to learning to read (National Institute of Child Health and Human Development, 2000). Specifically, they lead to reading through a process known as lexical restructuring, which begins with a child's growth in oral vocabulary and consequently leads to a heightened sensitivity to smaller and smaller units of sounds within words (see Anthony et al., 2009; Lonigan et al., 2013). This enables the emerging reader to proceed along a continuum of ever-increasingly sophisticated PA skills, beginning with detecting the blending and deletion of sounds, and later performing the blending or deletion of sounds, a developmental pattern which appears to be similar in both Spanish and English (Anthony et al., 2011). In this way, vocabulary development allows phonological awareness to grow (Lonigan, 2007). This reciprocal relationship establishes PA as more than just a foundational skill; it continues to improve as other early reading skills develop (Goldenberg et al., 2014; Majsterek & Ellenwood, 1995; Metsala & Walley, 1998; Perfetti, Beck, Bell, & Hughes, 1987).

The importance of a strong foundation in PA has been underscored by research that shows that PA skills in L1 correlate with those in L2, transferring across languages (Anthony et al., 2009; Comeau, Cormier, Grandmaison, & Lacroix, 1999; Lesaux & Siegel, 2003). In fact, whether PA skills are first developed in L1, L2, or both, a child's PA skills are a better predictor of reading ability in that language than the language of instruction (Leafstedt & Gerber, 2005). Further, because PA is language-general rather than language-specific (Branum-Martin et al., 2015), strong PA skills may help children develop reading and spelling skills in L2 (Geva & Siegel, 2000; Goodrich, Lonigan, & Farver, 2014; Lafrance & Gottardo, 2005). Other emergent literacy skills, such as vocabulary, may not transfer from L1 to L2 so readily (Goodrich et al., 2013).

A deficit in phonological awareness may predict future reading difficulties (Lonigan et al., 2013). Among ELs in particular, research has supported English PA as one of the strongest predictors of reading performance; however, low English oral language proficiency can cause an underestimation of reading abilities (see Lesaux & Siegel, 2003). The risk of interference of oral language decifits in the measurement of reading ability is a consistent challenge for reasearchers of ELs. Low English PA among ELs and its predictive validity for reading achievement among ELs is a key issue in the current study.

In summary, research shows that PA is a critical prerequisite to reading and develops along a continuum. Skills in PA not only transfer from L1 to L2 but also promote greater literacy skill development in L2. While PA's reciprocal relationship to reading skill development makes it important to all early readers, its transferability between alphabetic languages makes it particularly salient to Spanish-speaking ELs. Deficits in L2 (i.e., English) PA may suggest future reading difficulties in that language. An early awareness of deficits in PA, along with other important skill areas, such as oral language, may allow educators of ELs to better promote their development.

Theories of Oral Language Development

Like PA, oral language is a quantifiable and measurable skill that may be predictive of future reading abilities. Oral language is an important construct to consider alone because it develops differently from reading skills, though English oral language proficiency is important in the development of English reading skills for ELs (Lesaux & Geva, 2006). Oral language proficiency is defined by Lesaux and Geva (2006) as both receptive and expressive, encompassing knowledge or use of specific components, such

as phonology, vocabulary, grammar, and pragmatic skills. Saunders and O'Brien (2006) explain that research on oral language outcomes has consistently indicated that ELs require several years to develop English oral language proficiency, regardless of whether they are in a bilingual or immersion setting.

One important way in which oral language is different from other reading-related skills is that it does not appear to have an underlying process that enables oral language skills in L1 to readily transfer to L2 (Gottardo et al., 2011; Melby-Lervåg & Lervåg, 2011). The complex, multidimensionality of oral language may make transfer of its skills difficult to measure and therefore detect (Proctor et al., 2006). Its "psychometrically elusive" nature has led some researchers to theorize that there may be some transfer of oral language, for instance when languages share cognates (Proctor et al., 2006, p. 160), but most claim that there is no transfer of oral language skills (Cobo-Lewis, Eilers, Pearson, & Umbel, 2002; Gottardo et al., 2011; Manis & Lindsey, 2011). In contrast, phonological awareness has at its foundation the alphabetic principle-the knowledge that words can be broken down into smaller units-and acquisition of this fundamental insight facilitates the transfer of PA skills from L1 to L2 (Melby-Lervåg & Lervåg, 2011). This observed transferability is aligned with the linguistic interdependence theory proposed by Cummins (1979a).

There is a particularly important relationship between oral language skills and reading comprehension, which has implications for ELs' long-term academic achievement. Well-developed oral language skills are critical to reading comprehension, and so the achievement gap in reading comprehension between ELs and native speakers tends to be longer-lasting than gaps related to more discrete skills, such as word reading

or spelling (Lesaux & Geva, 2006). The connection between oral language and reading comprehension has been explained by the Simple View of Reading, which suggests that reading comprehension is predicted by decoding and oral language skills, as well as the interaction between the two (Gough & Tunmer, 1986; Mancilla-Martinez & Lesaux, 2010; Proctor et al., 2006). In fact, L2 oral language skills were found to have a stronger influence than word reading on reading comprehension (Lesaux, Crosson, Kieffer, & Pierce, 2010). So although ELs may be able to decode and read fluently without strong oral language skills (Durgunoğlu et al., 1993), they are not necessarily understanding what they read, and this is a key distinction, especially as students advance and are expected to learn from comprehending text.

Oral language development, in sum, plays an important role in reading development, particularly reading comprehension. It is a complex construct that is difficult to measure. There is some debate about the extent to which skills transfer to a second language, though it is understood that any transfer is limited. An assessment of oral language skills must be one component of any comprehensive assessment of an EL student's school functioning, and the proceeding paragraphs offer further guidelines for evaluating the presence of reading difficulties among EL students.

Best Practices in the Assessment of EL Reading Difficulties

When reading difficulties are suspected, the question of whether ELs are best assessed in their native language (L1) or their second language (L2), is a source of some debate. Beginning in 1997 and then reauthorized seven years later, the Individuals with Disabilities Education Improvement Act (IDEIA; 2004) requires that students be assessed in their native language whenever possible. Furthermore, without proper assessment,

school personnel may misunderstand, or worse misinterpret, a student's language abilities. Though basic interpersonal communication skills (BICS) may be strong, cognitive academic language proficiency (CALP), the more critical skill set for school success, may be underdeveloped (Cummins, 1979a, 2013). Additionally, second language acquisition takes time. Research shows that ELs who have received instruction in bilingual settings may take 4-7 years to perform at national norm levels on measures of English achievement (Hakuta et al., 2000; MacSwan & Pray, 2005). Relatedly, research supports the hypothesis that a certain level of competence, or a "linguistic threshold," must first be achieved in L1 for L2 to develop successfully, and that there is a "developmental interdependence" between competencies in both languages (Cummins, 1979b, 2009; Goodrich et al., 2013). The complexity of the evaluation process can be mitigated by measuring both L1 and L2 and is widely considered best practice because any true reading difficulty will be present in both languages (Cirino, Pollard-Durodola, Foorman, Carlson, & Francis, 2007; Townsend & Collins, 2008).

Given the importance of early intervention, a valid indicator of risk of reading difficulty is needed in both L1 and L2 for ELs. Measures of phonological awareness may meet that need. Early screening for PA difficulties may permit the identification of students at high risk of future reading difficulties to be identified early. It is well documented that PA screeners used early can predict reading difficulties in later years (Lonigan, 2006; Schatschneider, Fletcher, Francis, Carlson, & Foorman, 2004). Narrow-band measures that provide an assessment of critical reading prerequisites, such as PA, provide high predictive utility in identifying future reading difficulties (Lonigan, Schatschneider, & Westberg, 2008). A number of measures have been developed to

assess PA in emergent readers, including the Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgensen, & Rashotte, 1999), the Test of Phonological Awareness (TOPA; Torgesen & Bryant, 1995) or the Phonological Awareness subtest of the Kaufman Test of Educational Achievement, Third Edition (KTEA-3; Kaufman, Kaufman, & Breaux, 2014).

In sum, best practices in the assessment of ELs are established and include early screening for reading difficulties. Best practice requires assessment of ELs' abilities in both L1 and L2, and so early measurement of L1 and L2 PA skills may provide a gauge of risk of future reading difficulties and enable educators to intervene with instructional support during a critical window of opportunity. Measures of PA have been used to predict reading difficulties in emerging readers but assessing risk of reading difficulties among ELs is complicated by the entanglement of second language acquisition factors with typical development of reading ability. This is just one of many complex factors that render the instruction, evaluation, and study of ELs particularly difficult.

Difficulty in Researching EL Achievement

Gains have been made in understanding best practices in instructing and evaluating ELs, but several factors continue to pose challenges to researchers of EL student achievement performance. These include the inconsistent reporting of student outcomes across the country, the confounding effects of socioeconomic factors, the imprecision of learning disability identification, and, most importantly for the present study, the selection bias that affects student enrollment in different instructional programs.

Inconsistent reporting. Although NCLB (2002) requires states to report on the AYP of multiple subgroups, including ELs, it is left to the states to determine how best to identify, assess, and educate these students, resulting in vast inconsistencies across and even within states (Abedi, 2004). Such inconsistencies in tracking these students and their progress, as well as the varied instructional programs offered across the country, make comparing program outcomes difficult.

Inequity. Factors related to social, economic, and educational inequities further complicate the study of this population. Because so many ELs come from low-income families, socioeconomic status (SES) is often a confounding factor that must be considered (Hammer et al., 2014). Another variable that is difficult to disentangle from achievement data is the quality of instruction or the level of fidelity to the instructional program, since researchers often have little more than the school- or classroom-level categorical label to understand instruction type (Branum-Martin et al., 2015). Unfortunately, the instructional quality cannot be assumed adequate given that many students growing up in low-income communities are taught by underprepared teachers in schools with high staff turn-over (Calderón, Slavin, & Sánchez, 2011).

Disability identification procedures. Another challenge, though not unique to the EL population, is the lack of consensus regarding the most valid and reliable means by which students in need of academic intervention are identified. In research, when high-stakes decisions around eligibility for services are not occurring, a common practice for identifying students with a potential learning disability is to institute an arbitrary cut point at the 25th percentile: students who fall within the lowest quartile on a standardized measure of achievement are flagged as at-risk of or demonstrating a learning disability

(Francis, Fletcher, Shaywitz, Shaywitz, & Rourke, 1996). Given approximately 3.5% of the American school-age population receives services for a specific learning disability in school (Gargiulo, 2015), this is a conservative criterion and likely captures more students than necessary (Fletcher et al., 1994), erring on the side of the preferred Type II error. The ubiquity of this standard allows for findings to be compared across studies (Murphy, Mazzocco, Hanich, & Early, 2007), a benefit that must be weighed against its harms. Disadvantages include the loss of power by dichotomizing the continuous variable produced by most achievement testing (Branum-Martin, Fletcher, & Stuebing, 2013) and the instability of such groupings over time (Francis et al., 2005).

Selection bias. The final challenge, selection bias, is particularly relevant to the present study because this study was designed to remediate the consequences of selection bias. Shadish, Cook, and Campbell (2002, p. 512) define selection bias as "when selection results in differences in unit characteristics between conditions that may be related to outcome differences." Selection bias plagues nearly all researchers of EL academic outcomes who are concerned with language of instruction because of the non-random way that ELs enter and exit different instructional programs (Branum-Martin et al., 2015; Francis et al., 2006). In most of this research, the researcher is not able to randomize which EL receives which type of instruction, and students may receive one type of instruction over another for important reasons (described below) that may also directly influence their academic achievement.

Research suggests that several key demographic variables have a strong influence on a family's choice to enroll in bilingual programming. In California, for example, researchers studying the effects of Proposition 227 on ELs discovered that students

attending schools offering primary language instruction to less than half of their students tended to have three things in common: they had a higher (1) initial English oral language proficiency upon entering school, (2) socio-economic status, and (3) likelihood of being taught by a credentialed teacher (Parrish et al., 2006). This is evidence of the risk of selection bias in studying these students because the finding suggests that the population of students participating in immersion programming tends to reflect a particular student profile. Conversely, students with an at-risk profile for academic underachievement (i.e., lower English proficiency, SES, and access to credentialed teachers) were more likely to participate in bilingual programming, rendering selection bias a major barrier to understanding effects of instruction on academic outcomes.

In Texas, the other home state for participants in the current study, a gap in the literature exists. Perhaps because access to bilingual programming in Texas has been mandated by state law since 1981 and has yet to come under serious threat of elimination (K. V. Dixon, 2014), very little relevant research exists to help define the demographic profile of bilingual program participants in Texas. A further distinction between California and Texas is that Proposition 227 required that California parents actively seek out bilingual programming (Parrish et al., 2006), whereas program participation in Texas may have been a more passive choice, thus introducing yet another source of bias. That is to say, the choice to enroll in bilingual programming in Texas may be a function of what is readily available in one's neighborhood school. And although bilingual education in Texas may not be restrained by policy, human capital and economic realities have limited its growth (Bustos Flores, Keehn, & Perez, 2002). The Texas Education Agency has often reported on the shortage of highly qualified bilingual teachers, and school districts

may turn to waivers or long-term substitute teachers to address the shortfall, to the possible detriment of student learning (Kennedy, 2018). The sources of selection bias identified in the literature around bilingual program participation, therefore, include initial English oral language proficiency, family SES, and home state.

Beyond demographic variables, other more difficult-to-measure elements, such as beliefs about the benefits of bilingualism, may influence participation in bilingual education. Results from several studies of parent beliefs suggest that parents often enroll their children in bilingual education because they value bilingualism for its potential to yield future professional opportunities, cognitive benefits, and cultural competence/communication with family members (Giacchino-Baker & Piller, 2006; M. López, 2013; Shannon & Milian, 2002). Interestingly, parents of children in bilingual programming do not necessarily believe that this instructional approach is the most effective way to learn English. In California, when parents who had opted in to bilingual education (i.e., had requested waivers to participate) were surveyed about their educational beliefs, 20.7 percent of them reported that "submersion" was the "best way to acquire English" and just 46 percent endorsed "bilingual instruction" as best (Garcia, 2010). This suggests that some parents may have been influenced by factors beyond beliefs about pedagogy, such as convenience or preference for a particular teacher. Selection bias, therefore, is a complex challenge; a wide range of variables, some readily observable and others not, appear to influence a child's enrollment in bilingual education.

In summary, researchers of the essential question of how best to support the achievement of EL students are burdened by data-reporting limitations on the federal and state level. Additionally, confounding variables, particularly those tied to inequities and

the SES level of EL students and their communities, are difficult to take into account. Instituting a cut point for identifying students with learning disabilities in research creates statistical challenges but nonetheless remains a common standard, allowing for comparisons across studies. Finally, selection bias challenges researchers' ability to isolate the true impact of different instructional programs because of the non-random way that students participate in bilingual programming.

Previous Study

Given the aforementioned methodological challenges, and the broad concern for the risk of second language reading ability to be confounded with performance on oral language skill among ELs, long-term studies of ELs at risk for reading difficulties are critical to improving the process of early identification and intervention (Vaughn et al., 2008). In a previous study (Hilliard, 2016), a model of evaluation was proposed that combined L1 (Spanish) and L2 (English) measures of phonological awareness (PA) to identify Grade 1 EL students at risk for reading and oral language difficulties. Implementing the common practice of a cut point at the 25th percentile for demarcating low PA (Fletcher et al., 1994; Francis et al., 1996; Siegel & Ryan, 1989; Swanson, Sáez, & Gerber, 2006; Townsend & Collins, 2008), the model classified children into four groups: at-risk (low in both L1 and L2 measures), low Spanish PA (low in L1 but not low in L2), low English PA (not low in L1 but low in L2), and typical PA (not low in L1 or L2).

It was hypothesized that the lowest-performing group in phonological awareness in both languages (i.e., "at-risk" students) would go on to demonstrate reading or oral language difficulties when tested at the end of Grade 2, controlling for initial

performance at the beginning of Grade 1. Furthermore, it was hypothesized that assigned ability group based on PA scores would interact with students' instructional program, meaning that participation in bilingual versus immersion programming would yield different outcomes for different ability groups. Although the analyses yielded several significant results, there was no evidence of support for these original hypotheses.

Notable findings. Despite the lack of support for the original hypotheses, an interesting finding emerged regarding students with low PA in English: students with this profile benefitted most from bilingual instruction in reading outcomes in both languages. Although evidence indicates that the quality of instruction is more important than the instructional program in supporting positive student outcomes (Cheung & Slavin, 2012), this benefit for this particular group of students is supported by the theoretical basis of bilingual education that maintaining and further developing L1 will be beneficial to development in both languages (Genesee et al., 2004), and that early phonological awareness skills will confer cross-linguistic benefits in early reading skills (Anthony et al., 2009). The fact that oral language outcomes were no greater in bilingual education than immersion for this or other skill profiles is supported by the research reviewed previously: specifically, that oral language in L1 is greatly influenced by the home literacy environment, it takes several years to develop in L2 because it does not transfer from L1, and it is infrequently taught with the explicitness that EL students require. These findings taken together may indicate that the reading development of students with early signs of low English PA may be best supported in bilingual classrooms.

Limitations. Several limitations of the previous study must be acknowledged. First, researchers who collected the data from the sample had no way of randomly

assigning students to the immersion versus bilingual programming. Random assignment was impossible because students and their families controlled placement decisions (i.e., choosing to enroll in bilingual when the option of immersion was also available). Random assignment is a critical feature when determining causality between an intervention (i.e., instructional programming) and an outcome (i.e., reading or oral language measures) (Shadish et al., 2002). Without random assignment, selection bias (explained above) will surely interfere with the assessment of causality. Instead of random assignment, the previous study was an observational study, requiring that special precautions (further explained below) be taken to make reasonable inferences about the effect of the intervention (Rosenbaum, 2010). Methodological options to address limitations in the earlier study are outlined below.

Methodological Options

Due to the methodological challenges outlined above, a large focus of the current study was selecting the most appropriate methods given the unique features of the population in question. In consultation with an individual with expertise in such methods, it was determined that a thorough review of the literature around available methods was warranted. The proceeding paragraphs, though unconventional, are foundational to understanding the methodological decisions that were required to measure treatment effects between groups of students who self-selected into different programs, and therefore could not be randomized. After providing the context of the options available to the researcher, the methodological decisions are described in the Method section.

Randomized Controlled Trials: The "Gold Standard" and its Challenges

A randomized controlled trial (RCT), considered the "gold standard" of experimental research, hinges on the technique of random assignment, or a procedure that provides assurance that each participant has an equal chance of being assigned to each treatment condition to help control for the influence of extraneous variables and ensures that any pre-exiting differences are equally distributed (Gravetter & Wallnau, 2013). In the case of a randomized controlled trial, the treatment conditions are typically an experimental group that receives the treatment, and a control group that does not. The average treatment effect is equal to the average outcome of the treatment group minus the average outcome of the control group, with a standard error that can be estimated (Rubin, 2005). Effect sizes of the treatment may be assessed with less bias than observational trials because non-treatment variables should be equally influential to both groups, and any outcome differences can be attributed to the intervention with greater certainty (Sullivan, 2011). Such non-treatment variables include both measured and non-measured variables. Researchers may fail to account for non-measured variables because of an inability to measure them or a failure to recognize their existence. With a large enough sample, such error can be minimized (Deaton & Cartwright, 2018).

Employing the research method of randomization is difficult, especially in educational research. One of the primary obstacles is cost. According to a recent review of 56 published studies in the medical literature, the median cost per recruited participant was $409, with a cost range of $41-$6,990 (Speich et al., 2018). Another is acceptability. In addition to school personnel's general disinclination to withhold potentially helpful intervention from students who need it, other practical barriers are introduced regarding

the maintenance of treatment and control groups with the frequent movement of students and teachers between classes and schools, especially in high-needs academic settings (Berliner, 2002; Ritter & Maynard, 2008). ELs pose an additional challenge because of the non-random way that they enter into different instructional programs, and the wide range of teaching methods and quality of instruction that may be found across and within schools, districts, and states (Francis et al., 2006). Given the cost, the rigor, time, and control required to facilitate RCTs, it is not feasible to conduct RCTs to evaluate the effectiveness of the many interventions currently in use in applied settings, such as schools (Kazdin, 2010).

Observational Studies

Although randomization is the gold standard procedure for estimating causal effects, comparisons of nonrandomized groups (e.g., those enrolling in bilingual or immersion programming, as in this study) may be necessary to determine the effect of treatment when randomization is not feasible. Such comparisons may be misleading because of systematic differences between groups (Rosenbaum & Rubin, 1983), and it is these systematic differences that introduce selection bias, as described above. Procedures must be followed, therefore, that assist in accounting for systematic differences between treated and untreated subjects in terms of their baseline characteristics, or distinguishing features that existed between the groups before the treatment was introduced (Austin, 2011).

Demonstrating causality in observational studies is made difficult by systematic differences that may exist in the form of measured or unmeasured covariates. When control and treatment groups differ in either observable and/or unobservable ways, the

effect of treatment cannot be isolated and known without adequate precautions to control for such differences (Shadish et al., 2002). Findings from observational studies can be useful in themselves (Imai, King, & Stuart, 2008) but also as preliminary evidence to support the significant investment required for an RCT (Norman, 2003). Because researchers of ELs and language of instruction can rarely dictate how students enroll in school, most of this research has been observational in nature.

Methods for Controlling for Group Differences

Because systematic differences between groups in non-experimental research of treatment outcomes must be controlled, researchers have developed a number of methods to address the issue. The most common methods are described below.

Regression adjustment. One common method employed by researchers in education is regression analysis with nonrandomized data. Researchers regress independent variables such as pretest scores, demographics, and the treatment-related variables on post-test scores as a means for accounting for possible pretreatment effects (Pascarella, Wolniak, & Pierson, 2003). This method may help explain how treatment variables and outcomes are associated, but it is insufficient for explaining plausible causation (Grunwald & Mayhew, 2008) and the method relies on the proper specification of the relation between the treatment and the outcome (Kainz et al., 2017).

Matching. In order to create comparison groups for nonrandomized studies, matching techniques emerged (Rubin, 1973). A quasi-experimental matching procedure may be appropriate for assessing treatment effectiveness while reducing selection bias when randomization is not possible (Morgan, Frisco, Farkas, & Hibel, 2010). In this method, study participants are paired together based on their membership in one of two

naturally occurring treatment groups, such as those enrolled in two types of instructional programming (i.e., bilingual and immersion). For each individual in a treatment group, researchers attempt to find a close match in a comparison group based on a matching variable (Shadish et al., 2002). Due to its potential to improve accuracy in reporting results of treatment, the U.S. Department of Education (2003) has promoted matching as a means of establishing group equivalence in research when randomization is not possible.

Nevertheless, matching procedures are imperfect in yielding causal inference and many researchers of observational studies have offered recommendations for best practice (Austin, 2011; Lane, To, Shelley, & Henson, 2012; Thoemmes, 2012). In a discussion of the inherent shortcomings of a matching process, Shadish et al. (2002) explain that selection bias can never be fully eliminated because groups can never be matched to complete equivalence. In other words, units can only be matched on observed variables, so bias may still occur related to unobserved variables. Furthermore, matching proves to be difficult when variables are continuous or when there are many observed variables to match on (Grunwald & Mayhew, 2008).

Many researchers of observational studies have offered recommendations for best practice (Austin, 2011; Lane et al., 2012; Thoemmes, 2012). To improve the effectiveness of matching, Shadish and colleagues (2002) recommend selecting groups that are as similar as possible before matching, and matching on variables that are stable and reliable. Such selected covariates should be recognized in the research literature as contributing to the likelihood that a person will select the treatment condition, i.e., a person's propensity for treatment (Austin, 2011; Thoemmes, 2012).

Propensity scores. The propensity score method arose as an improvement on matching for simulating random assignment (Rosenbaum & Rubin, 1983). The propensity score is a variable useful in matching and was touted by Rosenbaum and Rubin (1983) as a tool that allowed for more meaningful comparison of nonrandomized treatment and control groups. A propensity score is defined as "the predicted probability of being in the treatment (versus control) group from a logistic regression equation" (Shadish et al., 2002, p. 162). Propensity scores can be used in four different ways to reduce the influence in confounds in assessing treatment outcomes: propensity score matching, stratification (or subclassification) on the propensity score, inverse probability of treatment weighting using the propensity score, and covariate adjustment using the propensity score (Austin, 2011). Given the relevance of propensity score matching to the current project, the section below explores in detail the issues and considerations related to this method.

Propensity Score Matching in Observational Studies

Because observational studies like the present do not allow for randomization, a quasi-experimental matching procedure may be appropriate for assessing treatment effectiveness while reducing selection bias (Morgan et al., 2010). In this method, study participants are paired together based on their membership in one of two naturally occurring treatment groups, such as those enrolled in two types of instructional programming (i.e., bilingual and immersion). In the present study, the method of propensity score matching (PSM) was chosen because of its ability to reduce the influence of selection bias in determining instructional program participation, thereby increasing the detectability of the treatment effects of the program.

The selection bias reduction that is possible through PSM has led to a proliferation of its use across disciplines, most notably epidemiology (Wang et al., 2018), public health (Li, Wen, & Henry, 2017), and economics (Lampach & Morawetz, 2016). It is a particularly useful method in education since randomly assigning students to various interventions is often difficult. Several recent studies of the effectiveness of educational interventions have made use of propensity score matching techniques (Dockx, De Fraine, & Vandecandelaere, 2019; Kretschmann, Vock, & Lüdtke, 2014; Kretschmann, Vock, Lüdtke, Jansen, & Gronostaj, 2019; Lane et al., 2012). Due to its potential to improve accuracy in reporting results of treatment, the U.S. Department of Education (2003) has promoted matching as a means of establishing group equivalence in research when randomization is not possible.

Developing the propensity score estimation model. An important early step in the PSM process is identifying the appropriate covariates to use in developing the propensity score estimation model. Often researchers recommend a theoretical approach, suggesting that selected covariates should be recognized in the research literature as contributing to the likelihood that a person will select the treatment condition (i.e., a person's propensity for treatment) (Caliendo & Kopeinig, 2008; Thoemmes, 2012). Only those variables unaffected by the treatment should be included in the model (Caliendo & Kopeinig, 2008). Stuart (2010) recommends being "liberal" with the inclusion of variables because omissions can lead to increased bias. Rubin and Thomas (1996) recommend excluding potential covariates only if there is a consensus in the literature that the covariate has no bearing on the outcome. Given the observed covariates, the

likelihood of a person participating in treatment is quantified as the propensity score (Caliendo & Kopeinig, 2008).

An iterative approach of model development using different combinations of baseline covariates is then followed to determine which combination yields the lowest standardized differences between the treated and untreated subjects (Austin, 2011), which is part of the process of assessing balance and discussed further below. The model development is typically completed with a series of logistic regression models (Lane et al., 2012), where group membership serves as the dependent variable and selected covariates serves as the independent variables (Grunwald & Mayhew, 2008). To improve balance in the model, non-linear terms (e.g., quadratic terms and interactions) may be included (Caliendo & Kopeinig, 2008). Linear model estimates may differ notably from non-linear estimates when there are outlier control units far outside the range of the treated units (Ho, Imai, King, & Stuart, 2017). Kainz and colleagues (2017) advise that any interaction that should be considered in a regression model should also be balanced in the matching model. When estimated through logistic regression, the propensity score is, as explained in plain language by Jacovidis and colleagues (2017, p. 536), "the probability of participation given the set of covariates. Moreover, students with the same propensity score are considered to have the same probability of participating in the intervention, regardless of whether or not they actually participated…[and outcomes between these two groups] can be compared conditioned upon the covariates included in the regression model."

Beyond the typical logistic regression approach, other ways that propensity scores are estimated include boosted regression, Bayesian regression, neural networks, and

classification and regression trees (see Austin, 2011; Ho et al., 2017; Jacovidis et al., 2017). Regardless of the estimation method, standard model diagnostics do not apply; in the case of estimating propensity scores, appropriate assessment of the model requires understanding the consequent balance of the covariates (Stuart, 2010). Steps involved in assessing balance are described further below.

Matching procedures. Choosing a matching procedure should include consideration of the nature of the sample data (e.g., sample size of the two groups, degree of overlap in the key covariates, and how they overlap). When beginning the matching process, one decision point is whether to include all cases or exclude outliers, or cases with propensity scores that do not fall within the overlapping distributions of the two groups (i.e., the common areas of support) (Stuart, 2010). Discarding cases outside of this region can improve balance (Thoemmes, 2012). If all cases are maintained, weighting methods or subclassification are strategies that may improve balance (Stuart, 2010).

Another decision to make is the ratio of control to treated cases among matches (e.g., one-to-one, many-to-one). One-to-one pair matching is the most common procedure (Austin, 2011). Some researchers have provided examples of how allowing for a variable number of control cases to be matched to a treatment case can allow for greater bias reduction (Ming & Rosenbaum, 2000). A matching procedure that permits one-to-many may be particularly helpful in improving statistical power when the sample sizes differ greatly among the two groups (Lane et al., 2012).

Choosing the most appropriate matching algorithm is another critical decision. This involves making a series of smaller decisions. Several authors have explained the options available (Austin, 2011; Caliendo & Kopeinig, 2008; Randolph, Falbe, Manuel,

& Balloun, 2014; Rosenbaum, 2010). Exact matching requires that matched pairs share the same propensity score (Thoemmes & Kim, 2011). Nearest neighbor, where a control case is matched to a treatment case based on similarity of propensity score, is a more common choice (Caliendo & Kopeinig, 2008; Randolph et al., 2014) and there are several options within it. There is the option to match with or without replacement. Matching with replacement may improve the quality of matching and reduce bias (Caliendo & Kopeinig, 2008) by requiring that after a control case has been used to make a match, it is still available to be matched with a different treated case (Austin, 2011). Alternatively, when cases are not replaced, then the independence of matched cases is preserved and additional steps such as frequency weighting are unnecessary (Stuart, 2010).

Another choice, optimal versus greedy matching, is discussed by Rosenbaum (2010). The greedy procedure, also known as the best-first algorithm, starts with a randomly selected treated case and matches it to the control case with the closest propensity score, even when that control case would be a better match with a different treated case. Optimal matching ensures that each pair has the closest possible set of propensity scores by resetting matches if a closer match is discovered. Caliendo and Kopeinig (2008) recommend randomly selecting the order in which cases are matched since order can affect the outcome of matching but do point out that optimal as opposed to greedy helps adjust for this. If multiple control cases are equally close to a treated case, then a control case is randomly selected to make the match (Austin, 2011).

As a means of controlling the quality of the matches, a limit can be set on the allowable distance between propensity scores among the matches cases. This limit is known as the caliper width, and it restricts the maximum distance that two matched cases

can be apart from each other in terms of their estimated propensity scores, thus protecting the balance of the matched pairs (Caliendo & Kopeinig, 2008). There is no consensus in the literature regarding a recommended allowable distance, but Rosenbaum (2010) recommends the caliper width be set to half of the standard deviation of the estimated propensity score.

In addition to nearest neighbor, another approach involves a stratification process that breaks up the overlapping area between groups into intervals and then compares them by calculating the mean difference in outcomes between treated and control cases (Luellen, Shadish, & Clark, 2005; Rosenbaum, 2010). This interval method may be a useful method when the sample size is large and there is a great deal of overlap between the treatment and control groups, or area of support (Grunwald & Mayhew, 2008).

Another alternative to nearest neighbor is kernel matching, which assigns a weight to each control case based on the closeness of its propensity score to the matching treated case propensity score (Morgan et al., 2010). All control cases may be used (Caliendo & Kopeinig, 2008). When more cases are included, efficiency increases but so does the risk for higher bias (Belfi, Haelermans, & De Fraine, 2016).

In summary, different matching procedures come with their own trade-offs in terms of efficiency and bias (Caliendo & Kopeinig, 2008). A decision must be made to account for the sample size and variation within the sample, and different methods should be tested for the balance they yield between groups. The process of assessing for balance is explained further below.

Evaluating the Quality of Matches

The quality of the matching procedure is measured by the degree of balance in the distribution of relevant covariates between the matched treatment and control groups. This balance assessment is a critical step in evaluating the propensity score estimation model; parameter estimates of the model are irrelevant (Stuart, 2010). Generally, the means or variances of covariates should be compared between groups before and after matching (Austin, 2011; Lane et al., 2012).

Rosenbaum and Rubin (1985) recommend two suitable methods for assessing balance. One of these is to determine the standardized difference in means of relevant covariates before and after matching, and Austin (2011) specifically recommends using pooled standard deviation to allow for comparison across units, unimpacted by sample size. The standardized difference is the difference in covariate means between the control and treated group divided by the pooled standard deviation (Kainz et al., 2017). A difficulty with this method is that there is no consistently-applied threshold for acceptability of match quality (Caliendo & Kopeinig, 2008); Austin (2011) cites research supporting a standard difference less than 0.1 standard deviations to suggest acceptable difference between groups on a covariate and Rubin (2001) recommends a standard mean difference of less than 0.2. Caliendo and Kopeinig (2008) have recommended more stringent criteria for sufficient balance, suggesting below 3-5% standard difference after matching.

The second method recommended by Rosenbaum and Rubin (1985) is to use a two-sample *t*-test to measure differences between groups' covariate means after matching. Adequate balance should yield no significant differences (Caliendo &

Kopeinig, 2008). Austin (2011) reviews reasons why this method may not be appropriate, chief among them being the fact that a matched sample is often reduced in size, thus limiting the ability of statistical significance testing to detect significant differences.

In order to assess the distribution of covariates between the groups, one recommended method is to produce a ratio of the variance of each variable (Kainz et al., 2017). Rubin (2007) proposes that optimal variance ratios would be between 0.8 and 1.2 to indicate good balance, but that a variance ratio between 0.5 and 2.0 is acceptable. Thoemmes (2011) simply recommended variance ratios close to 1.0.

Estimating Treatment Effects

Once the matched sample is created, the treatment effects can be estimated in much the same manner that they may be estimated in RCTs (Austin, 2011). According to Rosenbaum and Rubin (1983), assessing the difference between the mean outcome for control cases and the mean outcome for treated cases in the matched sample is adequate if the outcome variable is continuous. If the outcome is binary, they suggest that estimating the treatment effect can occur through estimating the difference between the proportion of cases that experience one of the two possible outcomes in each of the two groups. There is a wide range of options available for examining these differences after propensity score matching, from linear regression (Grunwald & Mayhew, 2008) to multilevel latent growth curve models (Dockx et al., 2019). There is some debate about whether the samples should be treated as paired or independent, as reviewed by Austin (2011). Ultimately, Austin argues that a propensity score matched sample is not comprised of independent observations, and recommends a paired approach with a variance estimator, such as a paired *t*-test.

Study Rationale

This study sought to minimize selection bias and re-examine the question of the relationship between instructional program and phonological awareness in terms of English and Spanish reading outcomes for ELs, building on a previous study (Hilliard, 2016). Taking into consideration the above literature review of available methods for minimizing selection bias, the current study employed the method of propensity score matching because it is a widely accepted way to isolate the treatment effect within the confines of an observational study when random assignment is not possible. As described above, it allows for balancing of baseline covariates between treatment and control groups thereby reducing bias introduced by covariates and allows for the estimation of causal effects.

Although this study began with a narrow focus on EL students with low English PA, it was not feasible to conduct an adequate matching procedure with such a small and disparate subsample. The revised study maintains a focus on the relationship between instructional programming, phonological awareness, and English and Spanish reading outcomes for EL students, but examines these relationships using the original sample with a full range of PA abilities.

Research Question and Hypothesis

The following research question and hypothesis were formed:

Research question. Do phonological awareness and instructional program interact to influence EL students' reading outcomes in English and Spanish?

Hypothesis. It was hypothesized that phonological awareness and instructional program would interact, and that the nature of that interaction would be that the lower the

phonological awareness, the greater the effects of bilingual instruction over immersion instruction on reading outcomes.in English and Spanish.

Chapter III:

Method

This study was conducted using data collected for a large, multisite, longitudinal project designed to explore the development of language and literacy skills among ELs in primary grades (see Branum-Martin et al., 2006; Cirino et al., 2007). A wide range of variables from the dataset was examined for balance between the instructional groups including reading, language and demographic variables. The reading and language variables are described below within the discussion of measures. Demographic variables are described below when discussing the sample.

Measures

All measures for the proposed study were administered to students first in Spanish and then in English, roughly one week apart, unless the child refused or was unable to master practice items for a test. Examiners were fluent in both Spanish and English, received standardized training in measure administration from the same trainers, and demonstrated mastery in administration through a certification process that involved conducting assessments under the supervision of the trainers. Data were collected from the beginning of kindergarten to the end of second grade, collected twice per year (i.e., beginning of year and end of year). For the current study, beginning-of-year first grade data served as pre-test data, and end-of-year second grade data served as post-test data.

Phonological awareness tests. Measures of phonological awareness, specifically the Comprehensive Test of Phonological Processing (CTOPP; Wagner et al., 1999) and its Spanish equivalent, the Test of Phonological Processing-Spanish (TOPPS; Francis et al., 2001), were administered at the beginning of Grade 1. The CTOPP is a well-validated

measure of phonological processing abilities with alpha coefficients for the subtests and age groups relevant to the current study ranging from .70 to .93. The TOPPS was designed to closely align with the CTOPP (i.e., not simply a translation of the CTOPP), measuring phonological processing skills in Spanish, and reliability tests with the present sample yielded alpha coefficients ranging from .93 to .97 (see Branum-Martin et al., 2006; Vaughn et al., 2006).Using items from the subtests (a) Blending Phonemes into Nonwords, (b) Blending Phonemes into Words, (c) Phoneme Elision, (d) Segmenting Words into Phonemes, and (e) First and Last Sound Identification, an item response theory (IRT) model was estimated to generate factor scores for phonological awareness in both languages for the beginning of Grade 1 (Goldenberg et al., 2014). The IRT model was based on scores from $N = 4{,}388$ U.S. bilingual and Mexican students, and the estimated mean (*SD*) for beginning of Grade 1 Spanish PA was -0.18 (0.88) (means and SDs for English PA were not provided). The Cronbach's alpha reliability coefficients for the subtests used to derive the factors scores were estimated to be greater than .92. In the present study, these factor scores were used to represent phonological awareness skills in each language.

Reading and oral language tests. Scores from the Woodcock Language Proficiency Battery-Revised (WLPB-R; Woodcock, 1991) and its Spanish form (Woodcock & Muñoz-Sandoval, 1995) were used to assess the change in reading skills in both English and Spanish from beginning-of-year Grade 1 to end-of-year Grade 2. The Oral Language cluster of the measure was also used as a broad-based measure of oral language proficiency in the generation of propensity scores (see below). The composite score of Basic Reading, composed of (a) Letter-Word Identification and (b) Word Attack,

was used to approximate reading abilities. The Oral Language cluster score, composed of subtest scores from (a) Memory for Sentences, (b) Picture Vocabulary, (c) Oral Vocabulary, (d) Listening Comprehension, and (e) Verbal Analogies, was used to approximate English oral language proficiency.

For both the Basic Reading composite and Oral Language cluster, *W* scores were reported, allowing for ease of comparison. The *W* scores are converted from raw scores based on a special transformation of the Rasch ability scale and are centered on a score of 500, considered to be the average score of a child age 10 years 0 months or at the start of Grade 5 (Woodcock, 1991). The *W* scores for the Basic Reading composite and Oral Language cluster are averages of the *W* scores from their respective subtests. Per WLPB-R norms for the start of Grade 1, the Basic Reading *W* score is 427 (*SEM* = 2) and the Oral Language *W* score is 474 (*SEM* = 2). In a review of reliability data for subtests for the Mental Measurements Yearbook, corrected split-half and internal consistency reliability coefficients were reported to be in the high .80s and low .90s (Lehmann & Poteat, 1995).

Participants

Sites and schools. The study sample included students enrolled in 35 schools across four different sites: Los Angeles, California, and Houston, Austin, and Brownsville, Texas. Selected schools served a population in which at least 40% of students identified as Hispanic and at least 30% of enrolled kindergarteners were designated as ELs. Using states' public school rating systems, researchers ensured that no failing schools were included in the study: selected Texas schools had a rating of at least "acceptable" and California schools met a comparable standard of a California Academic

Performance Index (API) score of 650. Though not required in the selection criteria, roughly 80% of families qualified for free and reduced lunch in most of the selected schools.

Program. Each selected school offered ELs either English immersion, dual language, early transition bilingual, or late transition bilingual. School principals reported the language program of their schools via survey and follow-up interviews with study personnel. For the purposes of the present study, and to ensure sufficiently large groups, early and late transition bilingual were collapsed into one group and treated as the bilingual instructional model, resulting in two distinct instructional programs: bilingual and immersion. Because so few students from dual language classrooms were included in the full sample, they were left out of the present study.

Students. The students for this study were drawn from the Oracy/Literacy Development of Spanish-Speaking Children project, and while this project has generated a great deal of research, Branum-Martin et al. (2006) were the first to publish a thorough explanation of the project's data collection methods. The initial matching sample consisted of 689 students (n = 427 in bilingual programs and n = 262 in immersion classrooms). The process through which this sample was established is visually depicted in Figure 1 and further explained below.

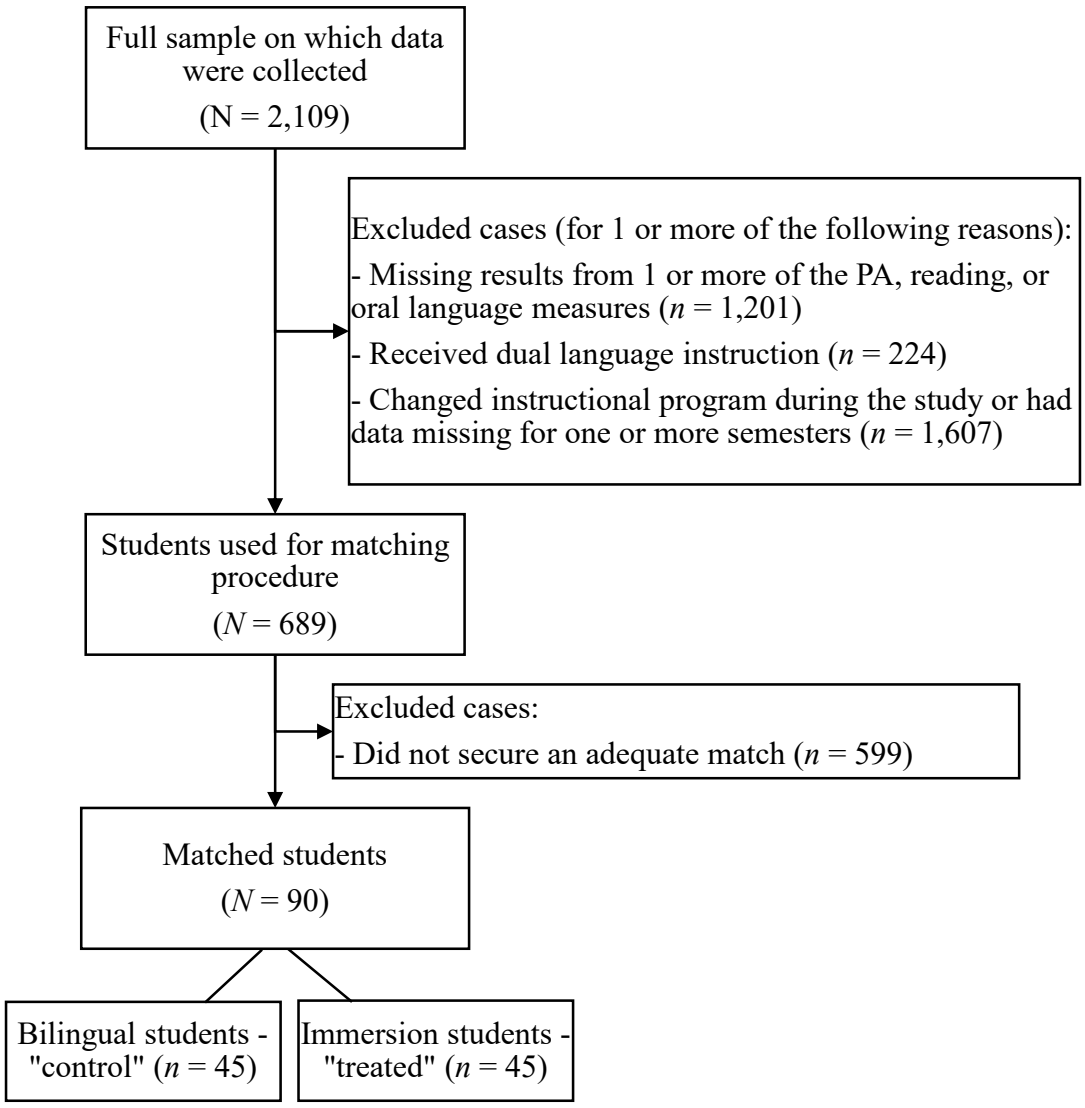

Figure 1. Flowchart depicting sample size changes

Of the 2,109 participants in the original study, 689 met the selection criteria for possible matched cases in the present study. Students were excluded (67% of original sample) from the eligibility pool if they (a) were missing results from one or more of the PA, reading, or oral language measures (n = 1,201), (b) received dual language instruction (n = 224), or (c) changed instructional programs during the study or had instructional program data missing for one or more semesters (n =1,607). Many students

were excluded based on more than one inclusion criterion. Students missing PA, reading, or oral language test results were excluded because these data were critical to the analysis in the present study. Those who received dual language instruction were too few in number to serve as a unique instructional program in the analysis. Lastly, students with changed or unreported instructional program information were considered to have received inconsistent curriculum and therefore could not be classified based on their instructional program.

The initial matching sample (n = 689) was used to generate propensity scores. Propensity scores should reflect a student's true propensity to enroll in bilingual education and will most accurately do so when based on as much data as possible (Rosenbaum & Rubin, 1983). These 689 students had the most stable instructional programming experience, in either bilingual or immersion settings, and therefore allowed for optimal generation of propensity scores. Additionally, these students had data available for all pertinent measures. Demographics of the sample are detailed in Table 1.

Table 1
Demographic Characteristics, Mean Scores, and Standard Mean Differences of Participants Before Matching

	Bilingual	Immersion	Standard Mean Difference	Test Statistic
n	427	262		
Female (%)	50.6	46.6	-0.08	1.050
Age – Mean months (SD)	79.01 (4.512)	77.69 (4.781)	-0.29	3.657***
Site (%)			-0.92	115.853***
California	24.6	66.0		
Texas	75.4	34.0		
Pre-Test English Phonological Awareness Factor Score	-.8988 (0.746)	-.9085 (0.784)	-0.01	0.163
Pre-Test Spanish Phonological Awareness Factor Score	-.9348 (0.869)	-1.0719 (0.744)	-0.17	2.120*
Pre-Test English Oral Language Proficiency	443.36 (14.255)	456.82 (12.086)	1.00	-12.739***
Pre-Test Spanish Oral Language Proficiency	469.72 (13.356)	452.0 (15.836)	-1.23	15.743***
Pre-Test English Reading	431.69 (24.046)	449.46 (16.715)	0.55	-10.506***
Pre-Test Spanish Reading	474.71 (36.523)	421.42 (31.021)	-1.54	19.664***

Note. Standard deviations are given in parentheses. Age is in months at start of First Grade. For sex and site, the test statistic was chi-square. For age and the pre-test continuous variables, the test statistic was a *t* test.
* $p < .05$. ** $p < .01$. *** $p < .001$.

As is shown in Table 1, bilingual and immersion groups significantly differed on most variables; however, for the purposes of creating matched samples, group differences are best explained in terms of standardized mean differences because statistical difference is greatly influenced by sample size. Standard mean differences (SMDs) indicated that

the greatest difference between bilingual and immersion groups prior to matching were on Spanish Oral Language (bilingual students were higher) and Spanish Reading (again, bilingual students were higher). The two groups were relatively close together in English PA, but immersion students were a full standard deviation higher in English oral language when compared to bilingual students. Immersion students demonstrated better English reading than their bilingual counterparts. Per national test norms, the average score for beginning-of-year Grade 1 in English and Spanish Oral Language was 474 (*SEM* = 2). For beginning-of-year Grade 1 in English and Spanish Basic Reading, the national norm was 427 (*SEM* = 3). Overall, the bilingual and immersion students were both lower in oral language than the typical first-grader. They were higher than the typical first-grader in reading except for the immersion students testing their Spanish reading skills.

Design and Analysis

The propensity score matched case design requires a set of procedures supported in the research literature as a way to mitigate the effects of bias in observational studies such as the present study. Initially, these procedures were to be guided by the exclusive use of the custom dialog PS Matching (Thoemmes, 2012) through IBM SPSS Statistics Version 26, which incorporates the Matchit program found in *R* software (version 3.6.1 used) by way of R Essentials. This system includes a comprehensive set of research-supported propensity score estimation and matching options and methods for evaluating the balance between groups; however, because of unique features of the data, the software alone was inadequate for creating relatively balanced groups with adequate sample sizes. The immersion and bilingual groups were so distinctly different in their

pre-test covariates that membership in the immersion treatment group could be predicted with high accuracy using pre-test reading and oral language skills. Consequently, the software produced a preponderance of 1's and 0's for propensity scores resulting in few matches. Overriding the software at certain points was therefore necessary to bypass this problem, but the alternative methods used were still consistent with recommendations found in the research literature. The overall process included: (1) conducting exploratory analyses to determine which combination of variables would produce the optimal propensity score model for decreasing group differences on pretest measures while maintaining an adequate sample size for evaluating the research question, (2) estimating propensity scores and creating initial matches across the entire sample using the variables identified above, (3) identifying matches to keep that fall within a specified criterion for minimizing propensity score differences within matched pairs, and (4) evaluating the matched samples to determine if relative balance was achieved while maintaining an adequate sample size.

Developing the Propensity Score Model. Variables chosen to include in the propensity score model should influence participation in the treatment but be unaffected by participation in the treatment (Caliendo & Kopeinig, 2008). The background characteristics that influence a child's enrollment in immersion versus bilingual education, per the literature review above, include initial English oral language proficiency, family SES, and home state. Data related to SES was collected via parent survey and was missing for many cases. As a result, no reliable variable representing SES was available for the PSM process. Therefore, these characteristics were represented by the following independent variables, respectively: beginning-of-year Grade 1 WLPB-R

English Oral Language cluster W score and site. Because Rubin and Thomas (1996) recommend excluding potential covariates only if there is a consensus in the literature that the covariate has no bearing on the outcome (i.e., end-of-year Grade 2 English and Spanish Reading in this case), sex, age at entering Grade 1, beginning-of-year Grade 1 WLPB-R Spanish Oral Language cluster W score, beginning-of-year Grade 1 WLPB-R English and Spanish Reading cluster W scores, and PA factor scores in English and Spanish were also considered as potential covariates. Therefore, a total of nine covariates were carefully considered for inclusion in the propensity score model.

Once the initial set of variables were identified for potential inclusion in the propensity score model, exploratory analyses were conducted to determine the final set of variables used for propensity score estimation. These analyses included examination of means and variances, histograms, and scatterplots to look for differences across treatment groups prior to matching. As described by Austin (2011), standard mean differences (SMD) using pooled standard deviations were generated for each covariate to assess the degree of similarity between the instructional groups. Groups are considered to be adequately similar on a covariate if the absolute value of their SMD is less than 0.2, per recommendation of Rubin (2001). This criterion was selected among those recommended (Austin, 2011; Caliendo & Kopeinig, 2008) as the least stringent but still acceptable level of similarity to support the matching process of two highly disparate groups. Variables that violated this criterion to the greatest degree were prioritized for potential inclusion in the final propensity score model.

Logistic regression models of different combinations of the variables described above were evaluated to determine the relative effects these combinations had on

predicting instructional group membership among the initial matching sample (n = 689). Logistic regression is the most commonly selected method for estimating propensity scores (Austin, 2011; Lane et al., 2012; Rosenbaum & Rubin, 1983). The PS Matching software uses logistic regression to estimate propensity scores; however, it does not provide the logistic regression parameters to help determine the best selection of variables to include in the final model, so this part of the procedure was done outside of the software. Once the optimal propensity score model incorporating a subset of the variables described above was identified, propensity scores were estimated for the entire matching sample. In the current study, the propensity score was interpreted as the probability of receiving immersion instruction.

Choice of matching algorithm and assessment of balance. There are several matching algorithms available to researchers. With a large sample size, all PSM methods will yield similar results; however, when a sample size is relatively small, as in the present study, multiple methods should be tested to determine which yields the best balance of the covariate distribution, including exact matching, kernel matching, and nearest neighbor matching in its various forms (e.g., optimal versus greedy matching, with and without caliper) (Caliendo & Kopeinig, 2008). The method of nearest neighbor one-to-one matching was ultimately selected for the present study, not only because it is the most common (Caliendo & Kopeinig, 2008; Randolph et al., 2014) and considered the simplest and most straight-forward matching method (Thoemmes & Kim, 2011), but also because it yielded the best balance in instructional groups after matching. Nearest neighbor matching selects for each treated case the control case whose propensity score is closest (Rubin, 1973). For technical reasons, students in the immersion group were

classified as treated cases and students in the bilingual group as control. The software selects each treated case one at a time and matches it to its closest control case (Thoemmes, 2012). Because there were more bilingual cases ($N = 427$) than immersion cases ($N = 262$), it was important to designate the immersion group as the treatment group such that immersion cases would be less likely to be discarded in the matching process, thereby preserving power as much as possible (Ho et al., 2017).

In addition to selecting a matching algorithm, there are several other options to consider in identifying the best matches for inclusion in the final matched sample. These options include: (1) whether or not to exclude some cases from the matching process, (2) one-to-one or many-to-one matching, (3) matching with or without replacement, (4) the order in which treated cases should be selected for matching to control cases, and (5) the criteria for including matched pairs in the final matched sample. The choices for each of these options and the rationales for those choices are described below.

First, the decision was made not to discard cases outside the common area of support (i.e., those that do not fall within the overlapping distributions of the two groups). When nearest neighbor is used, limiting matches to only those within the area of common support is not critical because matches are occurring based on proximity (unlike in other methods such as kernel matching) (Caliendo & Kopeinig, 2008), and maximizing potential good matches was the goal given the sample size. Second, a one-to-one (versus many-to-one) match ratio was selected to avoid incorporating frequency weighting in future analyses (Stuart, 2010) and also because group sizes were not adequate to accommodate a different ratio. Third, matching without replacement was chosen to preserve the independence of matched cases.

Fourth, the order in which treated (immersion) cases were chosen to match to control (bilingual) cases was from smallest to largest propensity score. This decision was based on the evaluation of three options: smallest first, largest first, or at random, to determine which yielded the best balance between groups overall. Although, in general, random selection of treated cases for matching to the closest control is recommended when using nearest neighbor (Caliendo & Kopeinig, 2008), the way EL students are selected into instructional programs results in unique characteristics of the two groups that make selecting treated cases from smallest-to-largest propensity score a better order for matching. In the case of Spanish-speaking ELs, students tend to self-select into immersion classes when they have higher English and lower Spanish oral language skills, while the reverse is true for students in bilingual classes (Parrish et al., 2006). Starting with an immersion student with the highest English oral language score, for example, could result in matching the student to a bilingual student with a lower English oral language score that would be better matched to an immersion student farther down in the distribution. Therefore, closer matches overall are attainable when beginning with the immersion group's smallest propensity scores (e.g., lowest likelihood of enrolling in immersion classrooms based on pre-enrollment English language and/or reading scores).

Fifth, once all immersion students were matched to the closest bilingual students within the larger sample, the dataset was pared down using a criterion of at most 0.25 difference in the propensity scores between any matched pair (i.e., for any matched pair, the probability of the cases participating in the immersion instructional program differed by no more than 0.25). The cutoff of 0.25 was chosen because it provided the best balance between attaining relatively matched groups with adequate sample size. This

criterion was applied by hand even though the PS Matching software allows for the designation of a caliper width, which restricts the maximum distance that two matched cases can be apart from each other in terms of their estimated propensity scores, thus protecting the balance of the matched pairs (Caliendo & Kopeinig, 2008). However, using a caliper removes the option to experiment with the matching order, as a caliper automatically causes each treated case to be matched to one control case that is randomly drawn out of all control cases within the caliper (Thoemmes, 2012).

Finally, all of the choices described above were made to identify a matched sample that minimized pre-test differences (i.e., achieve balance) between the immersion and bilingual groups while maintaining an adequate sample size for conducting group comparisons on post-test measures. The PS Matching software provides an array of research-supported univariate and multivariate tests of balance; however, these are automatically conducted on the full matched sample. Because the matched sample in the current study was pared down to only those pairs deemed adequately matched (i.e., equal to or less than 0.25 difference of propensity scores among a matched pair), the balance statistics generated by the software did not apply. Balance was evaluated using standardized mean differences (SMD) between the groups on the nine covariates identified above. The goal was to achieve as close to $|SMD| < .20$ as possible on all covariates while still maintaining a sample size adequately large for analysis of treatment effects.

Analysis Plan

A new dataset was created that included only the final matched sample ($n = 90$). Although a paired sample *t*-test is the recommended approach (Austin, 2011) when

groups are successfully balanced through matching and groups are to be compared on post-test outcomes, the present study required ANCOVA for two reasons. First, the treatment and control groups were not optimally balanced and, second, the research question was regarding the potential influences of phonological awareness on treatment effects. The benefits of ANCOVA in this case were that it could control for any group differences not resolved by matching and also it could test for interactions.

ANCOVA was used to test for an interaction between key beginning-of-year first grade variables (phonological awareness in English and Spanish) and treatment (bilingual or immersion) on end-of-year second grade reading achievement outcomes in English and Spanish on the pre-matched and matched datasets. This resulted in four ANCOVAs total. Due to the limited sample size, ANCOVAS were performed with variables within the same language and not across languages (e.g., English phonological awareness and English reading but not Spanish reading). A post-hoc power analysis was conducted and is reported below.

Chapter IV:

Results

In this section, results from best practices for implementing and evaluating the success of a matching procedure, as documented in the Method section, are reported. These include logistic regression to select proper covariates, estimating the propensity score, matching, evaluating balance, and assessing differences between the matched group and the larger sample. Next in the chapter are reported results from the ANCOVAS and a post-hoc power analysis.

Initial Evaluation of Logistic Regression Models to Identify Covariates

Prior to estimating the propensity scores, the predictive nature of covariates individually and combined was assessed through logistic regression. The propensity score estimation process ultimately focused solely on reading and language variables because the inclusion of other variables (e.g., site) yielded propensity scores of 1's and 0's, and this limitation is further explained below in the discussion section. As is documented in Table 2, Spanish reading was better at predicting instructional program than English reading (79% versus 65% classified correctly). Including both oral language variables with reading variables did not notably improve the predictive ability of the model above both reading variables alone (95% versus 94% classified correctly). The Wald tests did indicate initial reading and oral language skills in both languages did significantly predict group membership.

Table 2

Results of Binary Logistic Regression Analysis for Propensity Score Estimation Models

	Percent Classified Correctly	Nagel-kerke R^2	β	*SE* β	Wald	*df*	Odds ratio (OR)	95% CI for OR	
								Lower	Upper
Model 1	64.7	0.19							
English Reading*			-0.04	0	82.62	1	0.96	0.96	0.97
Model 2	78.7	0.48							
Spanish Reading*			0.04	0	172.6	1	1.04	1.04	1.05
Model 3	93.9	0.84							
English Reading*			-0.14	0.01	135.37	1	0.87	0.85	0.89
Spanish Reading*			0.1	0.01	149.09	1	1.1	1.09	1.12
Model 4	93.9	0.84							
English Reading*			-0.14	0.01	133.73	1	0.87	0.85	0.89
Spanish Reading*			0.1	0.01	142.46	1	1.1	1.09	1.12
Age			-0.01	0.04	0.05	1	0.99	0.92	1.07
Model 5	94.5	0.85							
English Reading*			-0.12	0.01	91.43	1	0.89	0.87	0.91
Spanish Reading*			0.09	0.01	104.5	1	1.09	1.07	1.11
English Oral Language*			-0.04	0.01	5.87	1	0.97	0.94	0.99
Spanish Oral Language*			0.03	0.02	4.4	1	1.03	1	1.06

Note. The dependent variable was instructional program with immersion as the target category (intervention) and bilingual as the reference category (control).
* $p < .05$.

Results of this early exploratory procedure guided the propensity score estimation model testing, beginning first with Spanish reading as the matching variable (referred to as "Model 2"). The second propensity score estimating model included both English and

Spanish reading ("Model 3"). The third propensity score estimating model included English and Spanish oral language ("Model 6"). Model 6 was tested as an alternative to Model 5 because, as explained above, adding both oral language variables to both reading variables did not notably improve the model.

Propensity Score Estimation and Matching

The three propensity score models (2, 3, and 6) were evaluated by estimating propensity scores, creating matched samples, and examining the balance (SMD's) between the immersion and bilingual groups across the nine variables described in the Method section. For each of the models, different matching orders (i.e., largest, random, and smallest) and criteria for including matched pairs (less than 0.20 versus 0.25 difference in propensity scores) were also evaluated. Results of this iterative, trial-and-error process of finding the best possible matching procedure yielding a sufficient number of matches and a well-balanced matched sample are detailed in Table 3. As shown in the table, matching in order of smallest first yielded more matches than largest or random. Also explained above, the criterion for an "adequate" match was originally set at a 0.2 difference in propensity scores but increased to 0.25 to increase the number of matches.

Table 3

Iterative Process to Identify Best Propensity Score Estimation Model

	Match Criteria of 0.20[a]		Match Criteria of 0.25[a]	
	Number of Adequate Matches	Average (range) of all SMDs for Covariates	Number of Adequate Matches	Average (range) of all SMDs for Covariates
Model 2				
Matching largest first	86	0.56 (0.06-1.47)		
Matching at random	127	0.61 (0.16-1.99)		
Matching smallest first	146	0.68 (0.04-2.47)		
Model 3				
Matching largest first	23	0.42 (0.05-0.72)		
Matching at random	32	0.25 (0.03-0.80)	35	0.24 (0.02-0.82)
Matching smallest first	36	0.22 (0.00-0.59)	45	0.19 (0.03-0.64)
Model 6				
Matching at random	102	0.35 (0.03-1.25)	114	0.37 (0.08-1.28)
Matching smallest first	118	0.35 (0.00-1.24)	119	0.35 (0.01-1.23)

Note. Model 2 is Spanish reading only, model 3 is English and Spanish reading, model 6 is English and Spanish Oral Language. Matching largest first was tested in Models 2 and 3, but consistently yielded the fewest matches, and so was not tested on Model 6. SMD is standard mean difference and is presented as an absolute value. Covariates included in the average of SMD's are: age, site, sex, pre-test reading (English and Spanish), pre-test oral language (English and Spanish), and pre-test phonological awareness factor scores (English and Spanish).

[a] Two levels of difference in propensity scores were used as matching criteria: 0.20 and 0.25

Model 3 was ultimately selected as the best model; it did not yield the most matches but it did yield the best balance with a sample size large enough to detect large effects. A detailed explanation of the thorough balance assessment is offered below, but a

simple representation of balance, quantified by the average of all covariates' standard mean differences between groups in the matched sample, is offered in Table 3.

Evaluating Balance

An analysis of group differences after matching helped to illustrate the applicability of the PSM process with this population, the adequacy of the matching, and the ways in which the sample changed after the matching procedure. Significance testing was conducted to detect any significant differences between groups, and chi-square and *t* test statistics, as appropriate given the nature of the variable. Standard mean differences between groups were compared before and after matching. Finally, improvements in variance ratios were evaluated.

Among the cases ultimately selected for the matched sample, there were 90 cases equally split between immersion and bilingual groups (see Figure 1). Table 4 reports means, standard deviations, standard mean differences (SMDs), and significance testing results for bilingual versus immersion groups in the matched sample. The only group difference that was statistically significant was for the site variable. As was the case prior to matching, there were more immersion students in California and more bilingual students in Texas, $\chi^2(1) = 8.46$, $p < .01$. All other variables were not significantly different between immersion and bilingual groups. The SMDs suggest that site retains the greatest group difference (-0.64), followed by the English (0.27) and Spanish oral language (-0.28). All other variables were within recommended criteria for achieving balance between the groups.

Table 4

Post-match Group Differences Between Treatment and Control Cases

	Treatment (Immersion) *M (SD)*	Control (Bilingual) *M (SD)*	*df*	Test Statistic[ab]
n	45	45		
Female (%)[a]	57.8	60	1	0.05
Age[b]	79.11 (4.62)	78.89 (4.08)	88	-.24
Site (%)			1	8.46**
California[a]	26.7	4.4		
Texas[a]	73.3	95.6		
English Phonological Awareness Factor Score[b]	-.55 (1.01)	-.59 (.69)	88	-.20
Spanish Phonological Awareness Factor Score[b]	-.73 (.95)	-.85 (9.82)	88	-.62
English Oral Language Proficiency[b]	453.51 (14.16)	449.60 (14.90)	88	-1.28
Spanish Oral Language Proficiency[b]	462.2 (12.64)	466.73 (19.00)	88	1.33
English Reading[b]	449.47 (23.33)	454.64 (29.71)	88	0.92
Spanish Reading[b]	457.93 (34.33)	458.89 (41.53)	88	0.12
Estimated Propensity Score	0.43 (0.23)	0.50 (0.31)	88	1.24

Note. Age is in months at start of First Grade.
[a]Chi-square tests were run for categorial variables. [b]*t* tests were run for continuous variables
* $p < .05$. ** $p < .01$. *** $p < .001$.

Next, standardized mean difference was assessed between treatment and control groups both before and after matching. Table 5 and Figure 2 illustrate the improvement in balance from pre-match to post-match through comparing standard mean differences before and after the matching procedure. Whereas prior to matching, six variables exceeded the recommended difference threshold of 0.2 (Rubin, 2001), after matching,

only English and Spanish Oral Language variables were inadequately balanced between groups.

Table 5

Covariate Balance Pre- and Post-Matching

	Pre-Match		Post-Match	
	Standard Mean Differences	Ratio: Immersion to Bilingual Variances	Standard Mean Differences	Ratio: Immersion to Bilingual Variances
Sex	-0.08	0.98	-0.04	1.02
Age	-0.29	1.12	0.05	1.27
Site	-0.92	1.21	-0.64	4.60
English PA	-0.01	1.10	0.04	2.11
Spanish PA	-0.17	0.73	0.13	1.35
English Oral Language	1.00	0.72	0.27	0.90
Spanish Oral Language	-1.23	1.41	-0.28	0.44
English Reading	0.55	0.48	0.19	1.62
Spanish Reading	-1.54	0.72	-0.03	0.68

Note: Variance ratios are the ratio of the variance of a variable in the immersion group to the variance of the variable in the bilingual group.

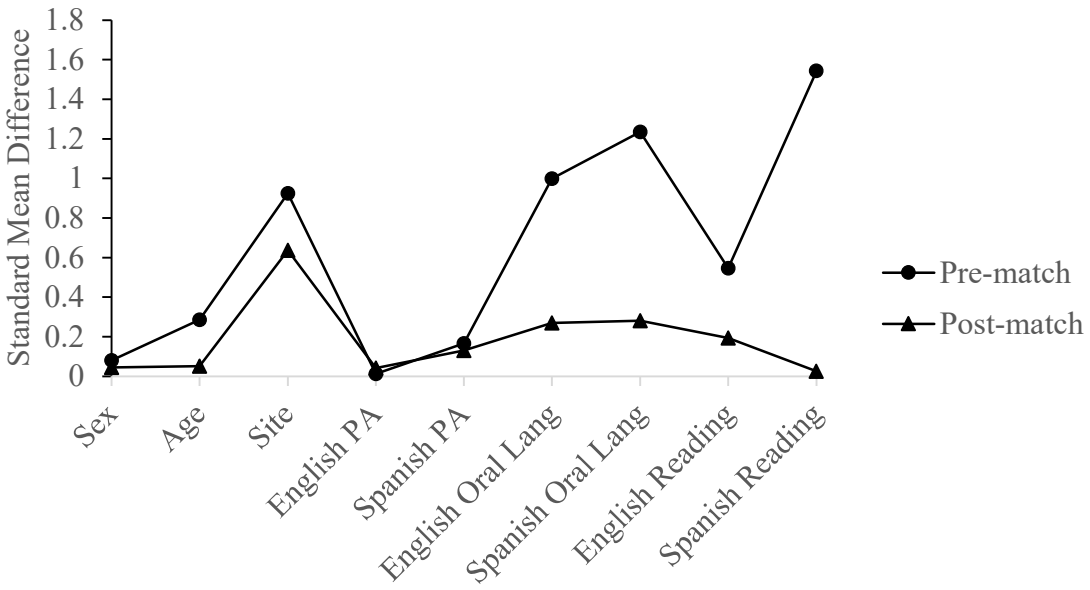

Figure 2. Absolute standard mean differences in the covariates between immersion and bilingual students pre- and post-matching. Values were calculated with pool standard deviations in the denominator.

An additional balance assessment involved examining the improvement in variance ratios, which should fall between 0.5 and 2.0 (Rubin, 2007) or be close to 1.0 (Thoemmes & Kim, 2011). Austin (2009) recommends calculating the variance of the variables to be compared between groups, and generating a ratio for each covariate before and after matching. Variances as well as variance ratios pre- and post-match are reported in Table 5. The variance ratios for the majority of covariates fell within the recommended limits as suggested by authors above. Exceptions included site, which saw even lower variance among the bilingual group once matched, and English PA, which increased in variance among the immersion group once matched.

Matched Sample as Representative of the Larger Sample of ELs

Matched students were included in the final analysis and unmatched students were excluded. Included and excluded students differed in important ways among immersion and bilingual students, which are fully documented in Tables 6 and 7, respectively. There was just one variable on which groups saw no significant difference in either immersion or bilingual groups; included (matched) cases did not differ significantly from excluded (unmatched) cases within immersion and bilingual groups on sex. Although not statistically significant, for both groups, females were more likely to be included than excluded (immersion: 58 versus 44%, bilingual: 60 versus 50%).

Group differences between included and excluded immersion students are detailed in Table 6. Among immersion students, there were significantly more cases matched from Texas than from California. Matched immersion students were significantly older than not-matched immersion students. Matched immersion students had significantly better English and Spanish PA scores. Matched immersion students had significantly worse English oral language and significantly better Spanish oral language. Matched immersion students had significantly better Spanish reading than not-matched immersion students. Matched immersion students had a significantly lower propensity score than those who were not matched. There was no significant difference between matched and not-matched immersion students in terms of English reading.

Table 6

Immersion Post-match Group Differences Between Included and Excluded Cases

	Included (Matched)	Excluded (Not Matched)		
	M (SD)	*M (SD)*	*df*	Test Statistic[ab]
n	45	217		
Female (%)[a]	57.8	44.2	1	2.745
Age[b]	79.11 (4.62)	77.39 (4.77)	260	-2.212*
Site (%)			1	37.533**
California[a]	26.7	74.2		
Texas[a]	73.3	25.8		
English Phonological Awareness Factor Score[b]	-0.56 (1.01)	-0.98 (0.71)	260	-3.394**
Spanish Phonological Awareness Factor Score[b]	-0.73 (0.96)	-1.14 (0.67)	260	-3.435**
English Oral Language Proficiency[b]	453.51 (14.16)	457.51 (11.53)	260	2.033*
Spanish Oral Language Proficiency[b]	462.20 (12.64)	449.88 (15.63)	260	-4.959***
English Reading[b]	449.47 (23.33)	449.46 (15.06)	260	-0.002
Spanish Reading[b]	457.93 (34.33)	413.84 (24.24)	260	-10.267***
Estimated Propensity Score	0.43 (0.22)	0.95 (0.06)	260	30.279***

Note. Age is in months at start of First Grade.
[a]Chi-square tests were run for categorial variables. [b]*t* tests were run for continuous variables
* $p < .05$. ** $p < .01$. *** $p < .001$.

Group differences between included and excluded bilingual students are detailed in Table 7. Among bilingual students, matched students were almost exclusively from Texas, and although the pre-match bilingual students were also majority-Texas, there was a significant difference between matched and not-matched bilingual students. Matched bilingual students were significantly better in their English PA, oral language, and reading. Matched bilingual students were significantly worse in their Spanish reading.

Matched bilingual students had a significantly higher estimated propensity score. There was no significant difference between bilingual groups in terms of age, Spanish PA, or Spanish oral language.

Table 7

Bilingual Post-match Group Differences Between Included and Excluded Cases

	Included (Matched)	Excluded (Not Matched)		
	M (SD)	*M (SD)*	*df*	Test Statistic[ab]
n	45	382		
Female (%)[a]	60.0	49.5	1	1.784
Age[b]	78.89 (4.10)	79.03 (4.56)	425	0.193
Site (%)			1	11.009**
California[a]	4.4	27.0		
Texas[a]	95.6	73.0		
English Phonological Awareness Factor Score[b]	-0.59 (0.69)	-0.94 (0.74)	425	-2.959**
Spanish Phonological Awareness Factor Score[b]	-0.85 (0.82)	-0.95 (0.88)	425	-0.711
English Oral Language Proficiency[b]	449.60 (14.90)	442.62 (14.02)	425	-3.139**
Spanish Oral Language Proficiency[b]	466.73 (19.0)	470.08 (12.51)	425	1.591
English Reading[b]	454.64 (29.72)	428.98 (21.79)	425	-7.159***
Spanish Reading[b]	458.89 (41.53)	476.58 (35.49)	425	3.104**
Estimated Propensity Score	0.50 (0.31)	0.03 (0.06)	425	-25.817***

Note. Age is in months at start of First Grade.
[a]Chi-square tests were run for categorial variables. [b]*t* tests were run for continuous variables
* $p < .05$. ** $p < .01$. *** $p < .001$.

In summary, on average, students with higher PA (English and Spanish) tended to be included in the matched samples across groups. In addition, the average English oral language shifted down for immersion and up for bilingual to attain a matched sample and

average Spanish oral language shifted in the opposite directions for the two groups (although the shift was not statistically significant for bilingual). English and Spanish reading shift in similar ways for the two groups with the exception of average English Reading among immersion students which did not shift at all to attain a matched sample.

Treatment Effects Estimation

For each test outcome, including reading in both English and Spanish, a one-way between-subjects covariance design was used to test for an interaction between phonological awareness and instructional program (bilingual or immersion). In each analysis, the dependent variable was the end-of-year Grade 2 reading test result. To statistically control for the degree to which students' starting point might affect performance, the beginning-of-year Grade 1 same-language reading test result was used as a covariate in the analysis. These ANCOVAS were run on pre-match sample but also on the matched sample in an attempt to mitigate the effects of selection bias. Results are reported in Table 8 and explained below.

Table 8

Analysis of Covariance of Reading Performance at End of Year Second Grade as a Function of Instructional Program and Phonological Awareness, with Reading Performance at Beginning of Year First Grade as Covariate

		Pre-Match Sample (n = 689)						Matched Sample (n = 90)					
		English			Spanish			English			Spanish		
Source	*df*	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2	*F*	*p*	η^2
Beginning of First Grade Reading (covariate)	1	293.40	<.001	0.30	408.52	<.001	0.37	50.66	<.001	0.38	61.35	<.001	0.42
Instructional Program (IP)	1	1.29	0.26	0.002	36.50	<.001	0.05	3.47	0.07	0.04	10.00	0.002	0.11
Phonological Awareness (PA)	1	1.20	0.27	0.002	7.81	0.01	0.01	2.36	0.13	0.03	0.81	0.37	0.01
IP x PA	1	0.20	0.66	0.000	2.74	0.09	0.004	2.42	0.12	0.03	0.32	0.57	0.004

English reading for pre-match sample. Figure 3 displays that, controlling for beginning-of-year Grade 1 English reading, bilingual students in the pre-match sample performed better on the end-of-year Grade 2 English reading test. The interaction between instructional program and phonological awareness was not significant, F (1, 683) = 0.20, p = 0.66. Groups did not significantly differ in terms of instructional program, F (1, 683) = 1.29, p = 0.26, or phonological awareness, F (1, 683) = 1.20, p = 0.27.

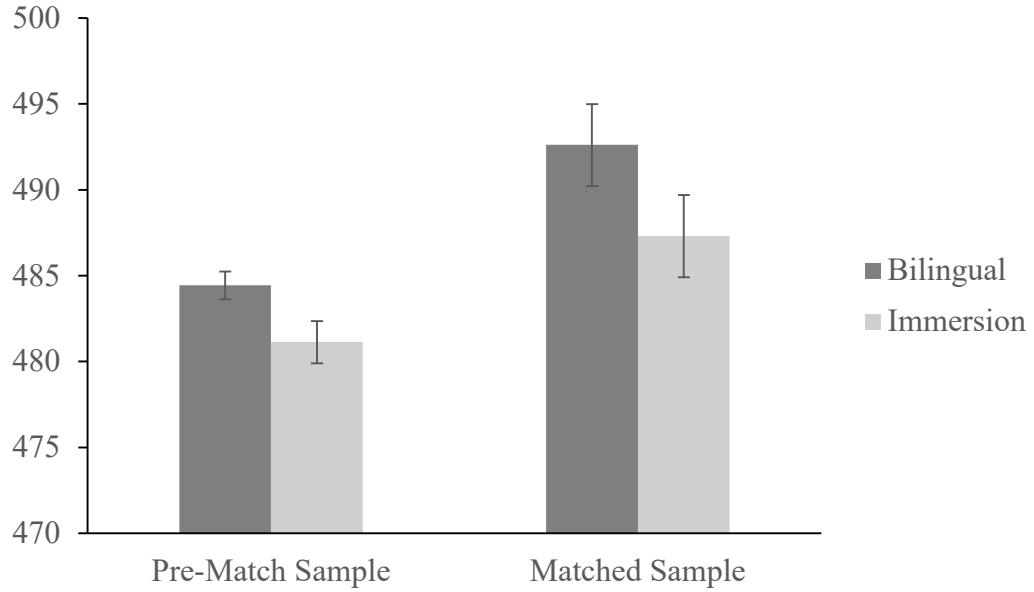

Figure 3. Marginal means for English Basic Reading *W* scores. These show end-of-year Grade 2 Basic Reading *W* scores in English while controlling for beginning-of-year Grade 1 Basic Reading *W* scores in English. Error bars show standard error.

English reading for matched sample. Similar to the pre-match sample above, Figure 3 displays that, controlling for beginning-of-year Grade 1 English reading, bilingual students in the matched sample performed better on the end-of-year Grade 2 English reading test. The interaction between instructional program and phonological

awareness was not significant, F (1, 84) = 2.42, p = 0.12. Again, groups did not significantly differ in terms of instructional program, F (1, 84) = 3.47, p = 0.07, or phonological awareness, F (1, 84) = 2.36, p = 0.13.

Spanish reading for pre-match sample. Figure 4 illustrates that, controlling for beginning-of-year Grade 1 Spanish reading, bilingual students in the pre-match sample performed better on the end-of-year Grade 2 Spanish reading test. The interaction between instructional program and phonological awareness was not significant, F (1, 683) = 2.74, p = 0.09. The ANCOVA yielded a significant main effect with the instructional program variable, F (1, 683) = 36.50, p < .001, partial η2 = 0.05, indicating that while controlling for beginning-of-year Grade 1 scores, students differed based on instructional program, with bilingual students outperforming immersion students. The ANCOVA also yielded a significant main effect with the phonological awareness variable, F (1, 683) =7.81, p = .01, partial η2 = 0.01, indicating that while controlling for beginning-of-year Grade 1 scores, students differed based on phonological awareness, with bilingual students outperforming immersion students.

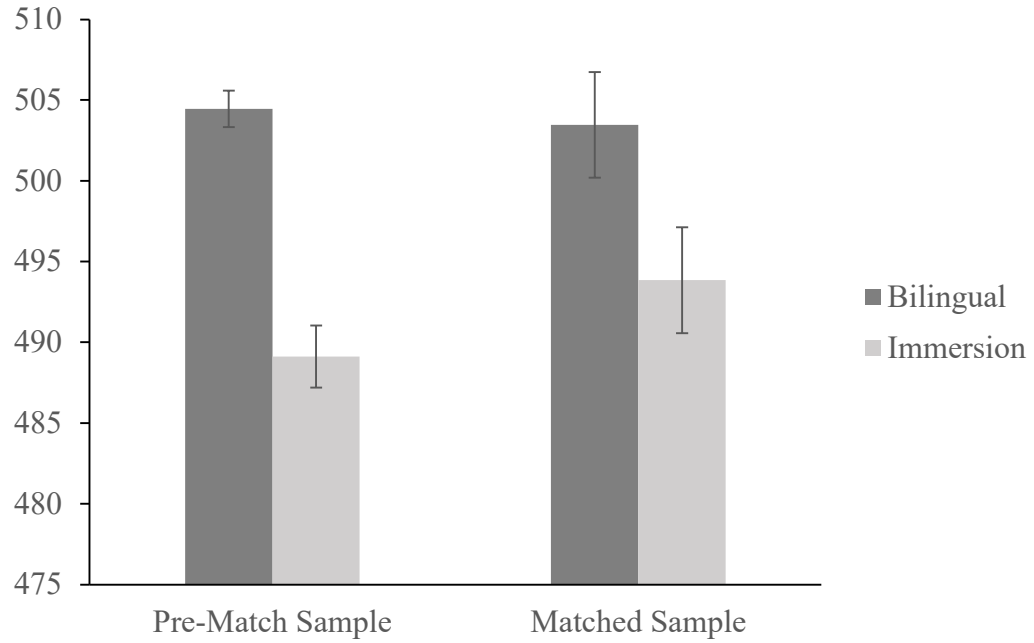

Figure 4. Marginal means for Spanish Basic Reading *W* scores. These show end-of-year Grade 2 Basic Reading *W* scores in Spanish while controlling for beginning-of-year Grade 1 Basic Reading *W* scores in Spanish. Error bars show standard error.

Spanish reading for matched sample. Similar to the pre-match sample above, Figure 4 displays that, controlling for beginning-of-year Grade 1 Spanish reading, bilingual students in the matched sample performed better on the end-of-year Grade 2 Spanish reading test. The interaction between instructional program and phonological awareness was not significant, F (1, 84) = 0.32, p = 0.57. The ANCOVA yielded a significant main effect with the instructional program variable, F (1, 84) = 10.00, p = .002, *partial* η^2 = 0.11, indicating that while controlling for beginning-of-year Grade 1 scores, students differed based on instructional program, with bilingual students outperforming immersion students. The ANCOVA did not yield a significant main effect with the phonological awareness variable, F (1, 84) = 0.81, p = 0.37.

Post-hoc Power Analysis

A post-hoc power analysis was conducted to determine whether the study was adequately powered to detect a treatment effect if one existed. The free software, G*Power3 (Faul, Erdfelder, Lang, & Buchner, 2007) was used to determine the power of the ANCOVAS for the interactions between phonological awareness and instructional program, and the main effect of instructional program on reading outcomes among the matched sample of 90 students with $\alpha = .05$. For the outcome of English reading, the power to detect an effect size of .18 for the interaction was .38 and the power to detect an effect size of .20 for the main effect was .48. For the outcome of Spanish reading, the power to detect an effect size of .06 for the interaction was .09 and the power to detect an effect size of .35 for the main effect was .91. Applying the standard minimum threshold of acceptable power at .80 (Shadish et al., 2002), the power analysis therefore demonstrated that only the ANCOVA for the main effect of instructional program on Spanish reading was adequately powered to detect an effect.

Chapter V:

Discussion

This study sought to examine whether phonological awareness and instructional program interact to influence EL students' reading outcomes in English and Spanish. Findings did not support the hypothesis that phonological awareness and instructional program would interact, and that the nature of that interaction would be that the lower the phonological awareness, the greater the effects of bilingual instruction over immersion instruction on reading outcomes in English and Spanish. There was no statistically significant interaction between instructional program and phonological awareness for either English or Spanish reading in the matched sample. This means that students' level of PA did not affect the level of impact that their instructional program had on their reading outcome, even after the matching procedure. This finding suggests that results fail to lend further support to the idea proposed by the author (2016) that results from screening students' early PA skills may help inform the selection of the most suitable instructional program for promoting the best possible reading outcomes. This idea was based on the previous finding that ELs with low English PA appeared to benefit more from bilingual education than immersion on English and Spanish reading outcomes.

There was a statistically significant main effect found for instructional program on Spanish reading outcomes in the matched sample. Specifically, students enrolled in bilingual classrooms performed significantly better in Spanish reading at the end of second grade, controlling for their Spanish reading ability at the start of first grade, compared to their counterparts in immersion classrooms. Based on the post-hoc power analysis for the matched sample, this ANCOVA was the only test adequately powered to

detect the effect. Furthermore, finding statistical significance on this outcome is logical for bilingual classrooms, where reading instruction is provided in Spanish, unlike immersion classrooms, which facilitate English-only instruction. As would be expected given the aforementioned goals of bilingual education, in which reading skill development in Spanish is a primary objective, bilingual instruction is likely to be the more effective option than immersion for developing Spanish reading. This is consistent with prior research, which has indicated that bilingual education has a positive effect on ELs' L1 reading (Nakamoto et al., 2012; Slavin & Cheung, 2005). This finding therefore suggests that families, educators, and policymakers who value Spanish reading skills should champion targeted Spanish reading instruction through bilingual education.

The lack of a significant interaction between instructional program and phonological awareness on Spanish reading may suggest that although bilingual students in this matched sample tended to outperform immersion students on Spanish reading outcomes, the nature of this relationship may not be affected in a consistent manner by their Spanish PA skill level (i.e., lower PA did not consistently increase the effect that bilingual education had on reading outcomes). Such a finding would be related to a previous finding by Goldenberg et al. (2014), which suggested that receipt of Spanish PA instruction had little bearing on ELs' long-term Spanish reading outcome. This suggests that when the goal of literacy instruction is strictly to promote Spanish reading skills, explicit PA instruction may not be needed; however, research on the transferability of PA between Spanish and English (Geva & Siegel, 2000; Goodrich et al., 2014; Lafrance & Gottardo, 2005) suggest that benefits of explicit PA instruction in Spanish may still exist in English reading outcomes. Therefore, the question of whether PA should be explicitly

taught and/or measured with screeners may be a function of the ultimate instructional goal.

Although not statistically significant, the English reading interaction and the instructional program main effect did have effect sizes approaching Cohen's (1992) standard of .25 for a medium effect with an *f* test. The effect size for the interaction was .18 and the effect size for the main effect was .20. Although there was still a high probability of committing a Type II error with these tests, these near-medium effect sizes are descriptively meaningful in that they suggest that the magnitude of the treatment effect was notable, though perhaps the sample size was not large enough to detect an effect with statistical significance. As is further described below, applying this method to this population necessitates a very large sample size because many cases will be excluded in an effort to achieve balanced groups. With a larger sample size, results from the ANCOVAS on the matched sample could have indicated a significant main effect for both English and Spanish reading outcomes in terms of instructional program, with bilingual students performing better than immersion students. Such a finding would have added support in favor of bilingual education, proponents of which have long argued that bilingual instruction is better for reading achievement in both L1 and L2 (August & Shanahan, 2006; Cummins, 2009; Goldenberg, 2013). A finding of better reading achievement in both languages by the end of second grade through bilingual education would present counterevidence to a major pro-immersion argument – that bilingual instruction hinders English reading growth (Rossell & Baker, 1996).

A lack of statistical differences in this study is likely a common occurrence for researchers of ELs and language of instruction, and this poses a problem for the field.

The influence of the unique features of this population (e.g., wide variation in Spanish/English oral language and pre-reading skills, as well as socio-political confounds such as immigration status, family SES, and access to high-quality education) and the important cultural, political, and economic forces (described previously) that influence their enrollment in different instructional programs, are challenging to control statistically. The inherent difficulty in facilitating randomized controlled trials in the school setting with an intervention as all-encompassing as language of instruction, for a sufficiently long period of time to measure meaningful outcomes, makes it a rare, though not impossible (August & Hakuta, 1997; Slavin et al., 2011), method. Most of the studies that met a high standard of methodological rigor and were included in a meta-analysis (Francis et al., 2006) of studies examining language of instruction and reading outcomes for ELs were designed to include a matching procedure, and many of those were unpublished dissertations. As was discovered in the current study, conducting matching procedures on such disparate groups of children (i.e., those enrolled in immersion versus bilingual) ultimately leads to a dramatically reduced sample size often not large enough to detect effects. This leads to bias in the field due to the "file drawer problem," and slows progress on addressing the question of optimal language of instruction for ELs (Francis et al., 2006).

Limitations

There were several major limitations that restricted the application of this method and the interpretation of its results. First, and most broadly, the selection bias that served as the original impetus for implementing a matching procedure to create balanced groups proved to be so extreme in this population that attempts at achieving balance came at

great cost to not only sample size but also how well the matched sample represented the larger sample and, by extension, the population. In this sample, compared to their counterparts in the other instructional program, bilingual students tended to have better Spanish phonological awareness, reading, and oral language, and immersion students tended to have better English phonological awareness, reading, and oral language. Students fell into these categories so neatly that attempts to match them on one variable created greater imbalance on another. These differences were in part due to the fact that, for the current study, baseline data were collected at the start of first grade, after students had completed kindergarten and received at least one year's worth of mostly-Spanish (for bilingual students) or English-only (for immersion students) instruction. This phenomenon is also aligned with the research literature, which suggests that students who have lower English language proficiency before beginning school are more likely to enroll in bilingual programming (Parrish et al., 2006).

Relatedly, the reading and oral language skills of students in the sample shifted upon implementation of the matching procedure, resulting in a limitation of generalizability to the broader EL population. The matched sample tended to exclude those bilingual students with stronger Spanish reading and immersion students with stronger English reading, as well as students across both groups with lower PA skills. Generally, this was because it was difficult to find adequate matches for students with relatively high or relatively low scores. Therefore, the matched sample was not only small, resulting in an underpowered study, but also restricted to those students with more moderate pretest skills (i.e., not high or low), rendering the sample less representative of the population.

The small sample size posed a limitation in terms of power. The initial groups differed greatly in size (e.g., $n = 427$ for bilingual and $n = 262$ for immersion), but the overall sample was not large enough to allow for one-to-many matching, which is the procedure that can help improve statistical power when the two groups differ in size (Lane et al., 2012). Conducting analysis on just 45 matched pairs limited the statistical power available for detecting differences between groups. Based on the power analysis, only one test was adequately powered to detect a treatment effect. Given the fact that some effect sizes approached the medium level, it is possible that a larger sample size would have provided power for statistical significance.

An additional limitation was that three important variables were not included as a control or matching variable. The first missing variable was socioeconomic status of the students. Children's language skills are closely associated with their parents' levels of income and education (Hart & Risley, 1995; Hoff, 2003) but, due to missing data in the original sample, socioeconomic status was not accounted for in either the matching procedures or the analysis. An additional variable excluded from the analysis was site, and as the historical and political review above suggests, geography greatly influences the type of instructional program that EL students receive. Precisely because of this great influence, a student's location in either Texas or California was so predictive of their instructional programming (i.e., Texas students were often bilingual and California students were often immersion) that it was not useful in the matching process. However, a possible consequence of not controlling for site differences could mean that the effect of instructional program is likely confounded by site differences (i.e., state learning standards, local curricula). A third important variable not included was any measure that

might quantify quality of instruction (e.g., instructional model adherence, evidence base of curriculum, years of experience of teachers), which has been touted repeatedly as perhaps the most important determinant in ELs' educational outcomes (L. Q. Dixon et al., 2012; Gersten & Baker, 2000; Slavin et al., 2011).

Future Directions

Theoretical. As documented in the literature review, American policy-makers, local education agencies, and families make choices about language of instruction for ELs based on a variety of factors, but without a firm consensus in the research literature about which method is the most beneficial to students. Researchers must continue to track longitudinal data on ELs' academic achievement through different instructional programs but also work to capture the data that represent the overall quality of the instruction provided through such metrics as time-by-activity and time-by-language, as others have done (Branum-Martin, Foorman, Francis, & Mehta, 2010; Foorman et al., 2006) and include in that data collection individual factors of ELs, such as not only phonological awareness skills but also other screenable early literacy skills like word reading or alphabetic knowledge. Such factors may predispose individuals to experiencing greater benefit from one type of programming over another, despite the lack of support for such an interaction in the current study. Research has shown that Spanish-speaking ELs, particularly those from low-income communities, enter school with wide variation in pre-literacy skills and that schools should be responsive to their needs with differentiated instruction (Gonzalez et al., 2015).

Methodological procedures for the current study highlighted the predictable way that ELs tended to enroll in either bilingual or immersion programs, and the literature

review explained some of the factors associated with those correlations. A culturally responsive approach to this research requires further investigation of the educational priorities and underlying motivations of ELs' parents and guardians, which are undoubtedly diverse. The question of what method is most effective for ELs at school is inextricably bound to the cultural values and educational beliefs held at home (Giacchino-Baker & Piller, 2006; M. López, 2013). A logical hypothesis would be that ELs would experience better school outcomes when their educational programming is aligned with the values reflected at home (e.g., bilingualism versus rapid mastery of English oral language) but this issue warrants further investigation as there is very little published work around it. Ultimately, families should have choices in the way that their children are educated, and those choices should be informed by both cultural values/educational beliefs and rigorous research.

Methodological. In order to apply the method of propensity score matching to ELs to detect differences between those receiving bilingual versus immersion instruction, the sample would have to be procured with great care. This population of students is unique in that the manner in which they enroll in different instructional programming is greatly influenced by not only personal factors, such as their own initial skillset, but also political and economic factors that influence the accessibility and desirability of those options.

Ideally, sampled students would have equal access to both bilingual and immersion options, so this choice would not be dictated by location. Robust background information, including family socioeconomic level, would be available for matching purposes. Pre-literacy skills levels, such as phonological awareness or oral language,

would be assessed prior to the start of instruction so as to avoid the influence of schooling. Furthermore, avoiding the possible confounds introduced by the varying learning standards and curricula dictated by site through limiting sampling to one geographic area would be ideal.

Because students with high Spanish skills opting in to bilingual classrooms and those with high English skills opting in to immersion classrooms is likely to be an unavoidable phenomenon, researches conducting propensity score matching with this population should consider procedural lessons learned from the present study. First, a very large sample size would allow for a one-to-many matching procedure, which could improve the quality of matches even with skillset-based sorting into classrooms. In particular, the "treatment" group should be over-sampled to maximize the potential to find adequate matches within that group for any given control case. Second, despite the general recommendation for a random order of matching propensity scores for nearest neighbor in most applications of this method (Caliendo & Kopeinig, 2008), an alternative approach may be better suited to this population. An extensive trial-and-error procedure with different matching procedures suggested that following a smallest-to-largest order of matching for ELs may provide the highest yield of adequately-matched pairs, as the distribution of students across literacy/language based variables between instructional settings is likely to be near-mirror images of each other.

Implications

This study highlights the need to take a more methodologically sound and rigorous approach to the complex and important work of evaluating reading outcomes for ELs engaged in different instructional programs. Researchers, policymakers, and

educators concerned with educational outcomes of ELs must consider the micro- and macrosystemic variables that influence the nature and quality of the instruction that ELs receive. Because of the highly predictable way that ELs sorted into bilingual and immersion classrooms in this sample, any matching method will likely require a very large sample to create balanced groups for comparison of treatment effects. Adequately long-term randomized controlled trials have been rare thus far but may be worth the investment in order to make informed decisions about what instructional approach works best for different Spanish-speaking ELs, a diverse group unified by their right to equitable educational opportunities in the United States. Their sizeable population now and projected growth in the future necessitate that instructional programming be guided not by what is politically palatable or readily available, but by converging evidence of effectiveness from high-quality research.

References

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher, 33*(1), 4-14. doi:10.3102/0013189x033001004

Anthony, J. L., Solari, E. J., Williams, J. M., Schoger, K. D., Zhang, Z., Branum-Martin, L., & Francis, D. J. (2009). Development of bilingual phonological awareness in Spanish-speaking English language learners: The roles of vocabulary, letter knowledge, and prior phonological awareness. *Scientific Studies of Reading, 13*(6), 535-564. doi:10.1080/10888430903034770

Anthony, J. L., Williams, J. M., Durán, L. K., Gillam, S. L., Liang, L., Aghara, R., . . . Landry, S. H. (2011). Spanish phonological awareness: Dimensionality and sequence of development during the preschool and kindergarten years. *Journal of Educational Psychology, 103*(4), 857-876. doi:10.1037/a0025024

Artiles, A. J., Rueda, R., Salazar, J. J., & Higareda, I. (2005). Within-group diversity in minority disproportionate representation: English language learners in urban school districts. *Exceptional Children, 71*(3), 283-300. doi:10.1177/001440290507100305

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

August, D., & Shanahan, T. (2006). *Developing literacy in second-language learners: Report of the National Literacy Panel on language-minority children and youth*. Mahwah, NY: Lawrence Erlbaum Associates, Publishers.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine, 28*(25), 3083-3107. doi:10.1002/sim.3697

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research, 46*(3), 399-424. doi:10.1080/00273171.2011.568786

Ball, E. W., & Blachman, B. A. (1991). Does phoneme awareness training in kindergarten make a difference in early word recognition and developmental spelling? *Reading Research Quarterly, 26*(1), 49-66. doi:10.1598/rrq.26.1.3

Belfi, B., Haelermans, C., & De Fraine, B. (2016). The long-term differential achievement effects of school socioeconomic composition in primary education: A propensity score matching approach. *British Journal of Educational Psychology, 86*(4), 501-525. doi:10.1111/bjep.12120

Berliner, D. C. (2002). Educational research: The hardest science of all. *Educational Researcher, 31*, 18-20. doi.org/10.3102/0013189x031008018

Bialystok, E. (2010). Global–local and trail-making tasks by monolingual and bilingual children: Beyond inhibition. *Developmental Psychology, 46*(1), 93-105. doi:10.1037/a0015466

Bialystok, E., & Shapero, D. (2005). Ambiguous benefits: The effect of bilingualism on reversing ambiguous figures. *Developmental Science, 8*(6), 595-604. doi:10.1111/j.1467-7687.2005.00451.x

Bilingual Education Act, Pub. L. No. 90-247, 81 Stat. 816 (1968).

Branum-Martin, L., Fletcher, J. M., & Stuebing, K. K. (2013). Classification and identification of reading and math disabilities: The special case of comorbidity. *Journal of Learning Disabilities, 46*(6), 490-499. doi:10.1177/0022219412468767

Branum-Martin, L., Mehta, P. D., Fletcher, J. M., Carlson, C. D., Ortiz, A., Carlo, M., & Francis, D. J. (2006). Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. *Journal of Educational Psychology, 98*(1), 170-181. doi:10.1037/0022-0663.98.1.170

Branum-Martin, L., Tao, S., & Garnaat, S. (2015). Bilingual phonological awareness: Reexamining the evidence for relations within and across languages. *Journal of Educational Psychology, 107*(1), 111-125. doi:10.1037/a0037149

Bustos Flores, B., Keehn, S., & Perez, B. (2002). Critical need for bilingual education teachers: The potentiality of Normalistas and paraprofessionals. *Bilingual Research Journal, 26*(3), 501-524. doi:10.1080/15235882.2002.10162575

Calderón, M., Slavin, R., & Sánchez, M. (2011). Effective instruction for English learners. *The Future of Children, 21*(1), 103-127. doi:10.1353/foc.2011.0007

Caliendo, M., & Kopeinig, S. (2008). Some practical guidance for the implementation of propensity score matching. *Journal of Economic Surveys, 22*(1), 31-72. doi:10.1111/j.1467-6419.2007.00527.x

Callahan, R. M. (2013). *The English learner dropout dilemma: Multiple risks and multiple resources* (Research Report No. 19). Retrieved from California Dropout Research Project website: http://www.cdrp.ucsb.edu/researchreport19.pdf

Capps, R., Fix, M. E., Murray, J., Ost, J., Passel, J., & Herwantoro-Hernandez, S. (2005). *The new demography of America's schools: Immigration and the No Child Left Behind Act*. Washington, DC: The Urban Institute.

Cárdenas-Hagan, E., Carlson, C. D., & Pollard-Durodola, S. D. (2007). The cross-linguistic transfer of early literacy skills: The role of initial L1 and L2 skills and language of instruction. *Language, Speech, and Hearing Services in Schools, 38*(3), 249-259. doi:10.1044/0161-1461(2007/026)

Cheung, A. C. K., & Slavin, R. E. (2012). Effective reading programs for Spanish-dominant English language learners (ELLs) in the elementary grades: A synthesis of research. *Review of Educational Research, 82*(4), 351-395. doi:10.3102/0034654312465472

Chiappe, P., Siegel, L. S., & Wade-Woolley, L. (2002). Linguistic diversity and the development of reading skills: A longitudinal study. *Scientific Studies of Reading, 6*(4), 369-400. doi:10.1207/S1532799XSSR0604_04

Christian, D. (1996). Two-way immersion education: Students learning through two languages. *Modern Language Journal, 80*(1), 66-76. doi:10.2307/329058

Christoffels, I. K., de Haan, A. M., Steenbergen, L., van den Wildenberg, W. P. M., & Colzato, L. S. (2015). Two is better than one: Bilingual education promotes the flexible mind. *Psychological Research, 79*(3), 371-379. doi:10.1007/s00426-014-0575-3

Cirino, P. T., Pollard-Durodola, S. D., Foorman, B. R., Carlson, C. D., & Francis, D. J. (2007). Teacher characteristics, classroom instruction, and student literacy and language outcomes in bilingual kindergartners. *The Elementary School Journal, 107*(4), 341-364. doi:10.1086/516668

Cirino, P. T., Vaughn, S., Linan-Thompson, S., Cardenas-Hagan, E., Fletcher, J. M., & Francis, D. J. (2009). One-year follow-up outcomes of Spanish and English interventions for English language learners at risk for reading problems. *American Educational Research Journal, 46*(3), 744-781. doi:10.3102/0002831208330214

Cobo-Lewis, A., Eilers, R. E., Pearson, B. Z., & Umbel, V. C. (2002). Interdependence of Spanish and English knowledge in language and literacy among bilingual children. In D. K. Oller & R. E. Eilers (Eds.), *Language and literacy in bilingual children* (pp. 118-132). Clevedon, UK: Multilingual Matters.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Comeau, L., Cormier, P., Grandmaison, É., & Lacroix, D. (1999). A longitudinal study of phonological processing skills in children learning to read in a second language. *Journal of Educational Psychology, 91*(1), 29-43. doi:10.1037/0022-0663.91.1.29

Cortez, A., & Johnson, R. L. (2008). Bilingual education in Texas – Where it is now, and what is still needed. *Intercultural Development Research Association Newsletter*.

Cummins, J. (1979a). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19*, 121-129.

Cummins, J. (1979b). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research, 49*(2), 222-251. doi:10.2307/1169960

Cummins, J. (2001). *Negotiating identities: Education for empowerment in a diverse society* (2nd ed.). Los Angeles, CA: California Association for Bilingual Education.

Cummins, J. (2009). Pedagogies of choice: Challenging coercive relations of power in classrooms and communities. *International Journal of Bilingual Education and Bilingualism, 12*(3), 261-271. doi:10.1080/13670050903003751

Cummins, J. (2013). BICS and CALP: Empirical support, theoretical status, and policy implications of a controversial distinction. In M. R. Hawkins (Ed.), *Framing languages and literacies: Socially situated views and perspectives* (pp. 10-23). London: Routledge.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine, 210*, 2-21. doi:10.1016/j.socscimed.2017.12.005

Dixon, K. V. (2014). *Framing bilingual education policy: Articulation and implementation in Texas*. Unpublished doctoral dissertation. University of North Texas, Denton, TX.

Dixon, L. Q., Zhao, J., Shin, J.-Y., Wu, S., Su, J.-H., Burgess-Brigham, R., . . . Snow, C. (2012). What we know about second language acquisition. *Review of Educational Research, 82*(1), 5-60. doi:10.3102/0034654311433587

Dockx, J., De Fraine, B., & Vandecandelaere, M. (2019). Does the track matter? A comparison of students' achievement in different tracks. *Journal of Educational Psychology, 111*(5), 827-846. doi:10.1037/edu0000305

Durgunoğlu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology, 85*(3), 453-465. doi:10.1037/0022-0663.85.3.453

Farver, J. M., Nakamoto, J., & Lonigan, C. J. (2007). Assessing preschoolers' emergent literacy skills in English and Spanish with the Get Ready to Read! screening tool. *Annals of Dyslexia, 57*(2), 161-178. doi:10.1007/s11881-007-0007-9

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. doi:10.3758/bf03193146

Fitzgerald, J. (1993). Views on bilingualism in the United States: A selective historical review. *Bilingual Research Journal, 17*(1, 2), 35-56.

Fletcher, J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., . . . Shaywitz, B. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. *Journal of Educational Psychology, 86*(1), 6-23. doi:10.1037/0022-0663.86.1.6

Foorman, B. R., Francis, D. J., Fletcher, J. M., Mehta, P., & Schatschneider, C. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology, 90*(1), 37-55. doi:10.1037/0022-0663.90.2.235

Foorman, B. R., Schatschneider, C., Eakin, M. N., Fletcher, J. M., Moats, L. C., & Francis, D. J. (2006). The impact of instructional practices in Grades 1 and 2 on reading and spelling achievement in high poverty schools. *Contemporary Educational Psychology, 31*(1), 1-29. doi:10.1016/j.cedpsych.2004.11.003

Francis, D. J., Carlo, M., August, D., Kenyon, D., Malabonga, V., Caglarcan, S., & Louguit, M. (2001). *Test of Phonological Processing in Spanish*. Washington, DC: Center for Applied Linguistics.

Francis, D. J., Fletcher, J. M., Shaywitz, B. A., Shaywitz, S. E., & Rourke, B. P. (1996). Defining learning and language disabilities: Conceptual and psychometric issues with the use of IQ tests. *Language, Speech, and Hearing Services in Schools, 27*(2), 132-143. doi:10.1044/0161-1461.2702.132

Francis, D. J., Fletcher, J. M., Stuebing, K. K., Lyon, G. R., Shaywitz, B. A., & Shaywitz, S. E. (2005). Psychometric approaches to the identification of LD: IQ and achievement scores are not sufficient. *Journal of Learning Disabilities, 38*(2), 98-108. doi:10.1177/00222194050380020101

Francis, D. J., Lesaux, N. K., & August, D. (2006). Language of instruction. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth.* (pp. 365-413). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Garcia, A. (2010). Informed parent consent and Proposition 227. *Bilingual Research Journal, 24*(1-2), 57-74. doi:10.1080/13825580701440510

Gargiulo, R. M. (2015). Individuals with learning disabilities *Special education in contemporary society: An introduction to exceptionality* (5th ed.). Los Angeles, CA: SAGE.

Genesee, F., Paradis, J., & Crago, M. B. (2004). *Dual language development and disorders: A handbook on bilingualism and second language learning*. Baltimore, MD: Paul H. Brookes Pub.

Gersten, R., & Baker, S. (2000). What we know about effective instructional practices for English-language learners. *Exceptional Children, 66*(4), 454-470. doi:10.1177/001440290006600402

Geva, E., & Siegel, L. S. (2000). Orthographic and cognitive factors in the concurrent development of basic reading skills in two languages. *Reading and Writing, 12*(1-2), 1-30. doi:10.1111/0023-8333.00087

Giacchino-Baker, R., & Piller, B. (2006). Parental motivation, attitudes, support, and commitment in a southern Californian two-way immersion program. *Journal of Latinos and Education, 5*(1), 5-28. doi:10.1207/s1532771xjle0501_2

Goldenberg, C. (2008). Teaching English language learners: What the research does-and does not-say. *American Educator, 32*(2), 8-44.

Goldenberg, C. (2013). Unlocking the research on English learners: What we know--and don't yet know--about effective instruction. *American Educator, 37*(2), 4-11.

Goldenberg, C., & Coleman, R. (2010). *Promoting academic achievement among English learners: A guide to the research*. Thousand Oaks, CA: Corwin.

Goldenberg, C., Tolar, T. D., Reese, L., Francis, D. J., Ray Bazán, A., & Mejía-Arauz, R. (2014). How important is teaching phonemic awareness to children learning to read in Spanish? *American Educational Research Journal, 51*(3), 604-633. doi:10.3102/0002831214529082

Gonzalez, J., Pollard-Durodola, S., Saenz, L., Soares, D., Davis, H., Resendez, N., & Zhu, L. (2015). Spanish and English early literacy profiles of preschool Latino English language learner children. *Early Education and Development, 27*(4), 513-531. doi:10.1080/10409289.2015.1077038

Goodrich, J. M., Lonigan, C. J., & Farver, J. M. (2013). Do early literacy skills in children's first language promote development of skills in their second language? An experimental evaluation of transfer. *Journal of Educational Psychology, 105*(2), 414-426. doi:10.1037/a0031780

Goodrich, J. M., Lonigan, C. J., & Farver, J. M. (2014). Children's expressive language skills and their impact on the relation between first- and second-language phonological awareness skills. *Scientific Studies of Reading, 18*(2), 114-129. doi:10.1080/10888438.2013.819355

Gottardo, A., Gu, Y., Mueller, J., Baciu, I., & Pauchulo, A. L. (2011). Factors affecting the relative relationships between first- and second-language phonological awareness and second-language reading. In A. Y. Durgunoglu & C. Goldenberg (Eds.), *Language and literacy development in bilingual settings* (pp. 141-167). New York: Guilford Press.

Gough, P. B., & Tunmer, W. E. (1986). Decoding, reading, and reading disability. *RASE: Remedial & Special Education, 7*(1), 6-10. doi:10.1177/074193258600700104

Graves, A. W., Gersten, R., & Haager, D. (2004). Literacy instruction in multiple-language first-grade classrooms: Linking student outcomes to observed instructional practice. *Learning Disabilities Research & Practice, 19*(4), 262-272. doi:10.1111/j.1540-5826.2004.00111.x

Gravetter, F. J., & Wallnau, L. B. (2013). Introduction to statistics *Statistics for the behavioral sciences* (9th ed., pp. 3-36). Bellmont, CA: Wadsworth.

Grunwald, H., & Mayhew, M. (2008). Using propensity scores for estimating causal effects: A study in the development of moral reasoning. *Research in Higher Education, 49*(8), 758-775. doi:10.1007/s11162-008-9103-x

Gyovai, L. K., Cartledge, G., Kourea, L., Yurick, A., & Gibson, L. (2009). Early reading intervention: Responding to the learning needs of young at-risk English language learners. *Learning Disability Quarterly, 32*(3), 143-162. doi:10.2307/27740365

Hakuta, K., Butler, Y. G., & Witt, D. (2000). *How long does it take English learners to attain proficiency*. Retrieved from http://www.escholarship.org/uc/item/13w7m06g

Halle, T., Hair, E., Wandner, L., McNamara, M., & Chien, N. (2012). Predictors and outcomes of early versus later English language proficiency among English language learners. *Early Childhood Research Quarterly, 27*(1), 1-20. doi:10.1016/j.ecresq.2011.07.004

Hammer, C. S., Hoff, E., Uchikoshi, Y., Gillanders, C., Castro, D. C., & Sandilos, L. E. (2014). The language and literacy development of young dual language learners: A critical review. *Early Childhood Research Quarterly, 29*(4), 715-733. doi:10.1016/j.ecresq.2014.05.008

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young American children*. Baltimore, MD: Brookes.

Hilliard, K. A. (2016). *Early prediction of reading difficulty among English language leaners*. Unpublished manuscript, Department of Psychological, Health, and Learning Sciences, University of Houston, Houston, TX.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2017). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis, 15*(3), 199-236. doi:10.1093/pan/mpl013

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development, 74*(5), 1368-1378. doi:10.1111/1467-8624.00612

Imai, K., King, G., & Stuart, E. (2008). Misunderstandings between experimentalists and observationalists about causal inference. *Journal of the Royal Statistical Society, 171*(2). doi:10.1111/j.1467-985X.2007.00527.x

Individuals With Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 (2004).

Jacovidis, J. N., Foelber, K. J., & Horst, S. J. (2017). The effect of propensity score matching method on the quantity and quality of matches. *Journal of Experimental Education, 85*(4), 535-558. doi:10.1080/00220973.2016.1250209

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*(4), 174-185. doi:10.1111/j.1540-5826.2009.00291.x

Kainz, K., Greifer, N., Givens, A., Swietek, K., Lombardi, B. M., Zietz, S., & Kohn, J. L. (2017). Improving causal inference: Recommendations for covariate selection and balance in propensity score methods. *Journal of the Society for Social Work and Research, 8*(2), 279-303. doi:10.1086/691464

Kaufman, A. S., Kaufman, N. L., & Breaux, K. C. (2014). Kaufman Test of Educational Achievement, Third Edition *KTEA-3*.

Kazdin, A. E. (2010). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kennedy, B. (2018). The bilingual teacher shortage in one Texas school district: Practitioner perspectives. *Journal of Latinos and Education,* 1-17. doi:10.1080/15348431.2018.1526688

Kretschmann, J., Vock, M., & Lüdtke, O. (2014). Acceleration in elementary school: Using propensity score matching to estimate the effects on academic achievement. *Journal of Educational Psychology, 106*(4), 1080-1095. doi:10.1037/a0036631

Kretschmann, J., Vock, M., Lüdtke, O., Jansen, M., & Gronostaj, A. (2019). Effects of grade retention on students' motivation: A longitudinal study over 3 years of secondary school. *Journal of Educational Psychology*. doi:10.1037/edu0000353

Kuo, L.-J., Uchikoshi, Y., Kim, T.-J., & Yang, X. (2016). Bilingualism and phonological awareness: Re-examining theories of cross-language transfer and structural sensitivity. *Contemporary Educational Psychology, 46*, 1-9. doi:10.1016/j.cedpsych.2016.03.002

LaCelle-Peterson, M. W., & Rivera, C. (1994). Is it real for all kids? A framework for equitable assessment policies for English language learners. *Harvard Educational Review, 64*(1), 55-75. doi:10.17763/haer.64.1.k3387733755817j7

Lafrance, A., & Gottardo, A. (2005). A longitudinal study of phonological processing skills and reading in bilingual children. *Applied Psycholinguistics, 26*(4), 559-578. doi:10.1017/s0142716405050307

Lampach, N., & Morawetz, U. B. (2016). Credibility of propensity score matching estimates. An example from Fair Trade certification of coffee producers. *Applied Economics, 48*(44), 4227-4237. doi:10.1080/00036846.2016.1153795

Lane, F. C., To, Y. M., Shelley, K., & Henson, R. K. (2012). An illustrative example of propensity score matching with education research. *Career and Technical Education Research, 37*(3), 187-212. doi:10.5328/cter37.3.187

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

Leafstedt, J. M., & Gerber, M. M. (2005). Crossover of phonological processing skills: A study of Spanish-speaking students in two instructional settings. *Remedial and Special Education, 26*(4), 226-235. doi:10.1177/07419325050260040501

Lehmann, I., & Poteat, G. M. (1995). Review of the Woodcock Language Proficiency Battery - Revised. In J. C. Conoley & J. C. Impara (Eds.), *The twelfth mental measurements yearbook*. Lincoln, NE: Buros Institute of Mental Measurements.

Lesaux, N. K. (2013). Reading and reading instruction for children from low-income and non-English-speaking households. *The Future of Children, 23*(2), 73-88. doi:10.1353/foc.2012.0010

Lesaux, N. K., Crosson, A. C., Kieffer, M. J., & Pierce, M. (2010). Uneven profiles: Language minority learners' word reading, vocabulary, and reading comprehension skills. *Journal of Applied Developmental Psychology, 31*(6), 475-483. doi:10.1016/j.appdev.2010.09.004

Lesaux, N. K., & Geva, E. (2006). Synthesis: Development of literacy in language-minority students. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth.* (pp. 53-74). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Lesaux, N. K., & Siegel, L. S. (2003). The development of reading in children who speak English as a second language. *Developmental Psychology, 39*(6), 1005-1019. doi:10.1037/0012-1649.39.6.1005

Lhamon, C. E., & Gupta, V. (2015, January 7). *Dear colleague letter: English learner students and limited Englsh proficient parents*. Retrieved from https://www2.ed.gov/about/offices/list/ocr/ellresources.html

Li, K., Wen, M., & Henry, K. A. (2017). Ethnic density, immigrant enclaves, and Latino health risks: A propensity score matching approach. *Social Science & Medicine, 189*, 44-52. doi:10.1016/j.socscimed.2017.07.019

Lonigan, C. J. (2006). Development, assessment, and promotion of preliteracy skills. *Early Education and Development, 17*(1), 91-114. doi:10.1207/s15566935eed1701_5

Lonigan, C. J. (2007). Vocabulary development and the development of phonological awareness skills in preschool children. In R. K. Wagner, A. E. Muse, & K. R. Tannenbaum (Eds.), *Vocabulary acquisition: Implications for reading comprehension* (pp. 15-51). New York: Guilford Press.

Lonigan, C. J., Farver, J. M., Nakamoto, J., & Eppe, S. (2013). Developmental trajectories of preschool early literacy skills: A comparison of language-minority and monolingual-English children. *Developmental Psychology, 49*(10), 1943-1957. doi:10.1037/a0031408

Lonigan, C. J., Schatschneider, C., & Westberg, L. (2008). Impact of code-focused interventions on young children's early literacy skills *Developing early literacy: Report of the National Early Literacy Panel* (pp. 107-151). Washington, DC: National Institute for Literacy.

López, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of Hispanic Head Start children. *Bilingual Research Journal, 28*(1), 1-18. doi:10.1080/15235882.2004.10162609

López, M. (2013). Mothers choose: Reasons for enrolling their children in a two-way immersion program. *Bilingual Research Journal, 36*(2), 208-227. doi:10.1080/15235882.2013.818595

Luellen, J. K., Shadish, W. R., & Clark, M. H. (2005). Propensity Scores: An Introduction and Experimental Test. *Evaluation Review, 29*(6), 530-558. doi:10.1177/0193841X05275596

MacSwan, J., & Pray, I. (2005). Learning English bilingually: Age of onset of exposure and rate of acquisition among English language learners in a bilingual education program. *Bilingual Research Journal, 29*(3), 653-678. doi:10.1080/15235882.2005.10162857

Majsterek, D. J., & Ellenwood, A. E. (1995). Phonological awareness and beginning reading: Evaluation of a school-based screening procedure. *Journal of Learning Disabilities, 28*(7), 449-456. doi:10.1177/002221949502800708

Mancilla-Martinez, J., & Lesaux, N. K. (2010). Predictors of reading comprehension for struggling readers: The case of Spanish-speaking language minority learners. *Journal of Educational Psychology, 102*(3), 701-711. doi:10.1037/a0019135

Manis, F. R., & Lindsey, K. A. (2011). Cognitive and oral language contributors to reading disabilities in Spanish-English bilinguals. In A. Y. Durgunoglu & C. Goldenberg (Eds.), *Language and literacy development in bilingual settings* (pp. 280-303). New York: Guilford Press.

McBride, H. E. A., & Siegel, L. S. (1997). Learning disabilities and adolescent suicide. *Journal of Learning Disabilities, 30*(6), 652-659. doi:10.1177/002221949703000609

Melby-Lervåg, M., & Lervåg, A. (2011). Cross-linguistic transfer of oral language, decoding, phonological awareness and reading comprehension: A meta-analysis of the correlational evidence. *Journal of Research in Reading, 34*(1), 114-135. doi:10.1111/j.1467-9817.2010.01477.x

Metsala, J. L., & Walley, A. C. (1998). Spoken vocabulary growth and the segmental restructuring of lexical representation: Precursors to phonemic awareness and early reading ability. In J. L. Metsala & L. C. Ehri (Eds.), *Word recognition in beginning literacy* (pp. 89-120). Mahwah, NJ: Erlbaum.

Ming, K., & Rosenbaum, P. R. (2000). Substantial gains in bias reduction from matching with a variable number of controls. *Biometrics, 56*(1), 118-124. doi:10.1111/j.0006-341X.2000.00118.x

Morgan, P. L., Frisco, M. L., Farkas, G., & Hibel, J. (2010). A propensity score matching analysis of the effects of special education services. *The Journal of Special Education, 43*, 236-254. doi:10.1177/0022466908323007

Motel, S., & Patten, E. (2012). *Characteristics of the 60 Largest Metropolitan Areas by Hispanic Population*. Retrieved from Washington, DC: http://www.pewhispanic.org/2012/09/19/characteristics-of-the-60-largest-metropolitan-areas-by-hispanic-population/

Murphy, M. M., Mazzocco, M. M. M., Hanich, L. B., & Early, M. C. (2007). Cognitive characteristics of children with mathematics learning disability (MLD) vary as a function of the cutoff criterion used to define MLD. *Journal of Learning Disabilities, 40*(5), 458-478. doi:10.1177/00222194070400050901

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2012). Development of reading skills from K-3 in Spanish-speaking English language learners following three programs of instruction. *Reading and Writing, 25*(2), 537-567. doi:10.1007/s11145-010-9285-4

National Academies of Sciences, Engineering, and Medicine. (2017). Programs for English learners in grades pre-k to 12. In R. Takanishi & S. Le Menestrel (Eds.), *Promoting the educational success of children and youth learning English: Promising futures*. Washington, DC: The National Academies Press.

National Assessment of Educational Progress. (2019). *The nation's report card*. Retrieved from: https://www.nationsreportcard.gov/#

National Center for Education Statistics. (2000). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Rockville, MD: National Institute of Child Health and Human Development Clearinghouse. Retrieved from https://www.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf.

National Center for Education Statistics. (2017). *Status and trends in the education of racial and ethnic groups*. Retrieved from https://nces.ed.gov/programs/raceindicators/indicator_raa.asp

National Center for Education Statistics. (2018). National student group scores and score gaps. Retrieved from https://www.nationsreportcard.gov/reading_2017/#/nation/gaps?grade=4

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching children to read. An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Part I: Phonemic awareness instruction*. Retrieved from https://www1.nichd.nih.gov/publications/pubs/nrp/Documents/report.pdf

Norman, G. (2003). RCT = results confounded and trivial: The perils of grand educational experiments. *Medical Education, 37*(7), 582-584. doi:10.1046/j.1365-2923.2003.01586.x

Ochoa, S. H. (2005). Bilingual education and second-language acquisition: Implications for assessment and school-based practice. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students: A practical guide* (pp. 57-75). New York: Guilford Press.

Olson, K. (2007). Lost opportunities to learn: The effects of education policy on primary language instruction for English learners. *Linguistics and Education, 18*(2), 121-141. doi:10.1016/j.linged.2007.07.001

Ortiz, C. J., Valerio, M. A., & Lopez, K. (2012). Trends in Hispanic academic achievement: Where do we go from here? *Journal of Hispanic Higher Education, 11*(2), 136-148. doi:10.1177/1538192712437935

Paradis, J., Genesee, F., & Crago, M. B. (2011a). The language-cognition connection. In *Dual language development and disorders* (2nd ed., pp. 49-53). Baltimore, MD: Paul H. Brookes Publishing Co.

Paradis, J., Genesee, F., & Crago, M. B. (2011b). Reading impairment in dual language children. In *Dual language development and disorders* (2nd ed., pp. 234-261). Baltimore, MD: Paul H. Brookes Publishing Co.

Parrish, T., Merickel, A., Pérez, M., Linquanti, R., Socías, M., Spain, A., . . . DeLancey, D. (2006). *Effects of the implementation of Proposition 227 on the education of English learners, K-12; Findings from a five-year evaluation*. Retrieved from https://www.wested.org/online_pubs/227Reportb.pdf

Pascarella, E., Wolniak, G., & Pierson, C. (2003). Explaining student growth in college when you don't think you are. *Journal of College Student Development, 44*(1), 122. doi:10.1353/csd.2003.0007

Passel, J. S., & D'Vera, C. (2008). *U.S. Population Projections: 2005 - 2050*. Retrieved from Pew Research Center website: http://pewhispanic.org/files/reports/85.pdf

Perfetti, C. A., Beck, I., Bell, L. C., & Hughes, C. (1987). Phonemic knowledge and learning to read are reciprocal: A longitudinal study of first grade children. *Merrill-Palmer Quarterly, 33*(3), 283-319.

Proctor, C. P., August, D., Carlo, M. S., & Snow, C. (2006). The intriguing role of Spanish language vocabulary knowledge in predicting English reading comprehension. *Journal of Educational Psychology, 98*(1), 159-169. doi:10.1037/0022-0663.98.1.159

Randolph, J. J., Falbe, K., Manuel, A. K., & Balloun, J. L. (2014). A step-by-step guide to propensity score matching in R. *Practical Assessment, Research & Evaluation, 19*(18).

Ritter, G., & Maynard, R. (2008). Using the right design to get the 'wrong' answer? Results of a random assignment evaluation of a volunteer tutoring programme. *Journal of Children's Services, 3*(2), 4-16. doi:10.1108/17466660200800008

Rolstad, K., Mahoney, K., & Glass, G. V. (2005). The big picture: A meta-analysis of program effectiveness research on English language learners. *Educational Policy, 19*, 572-594. doi:10.1177/0895904805278067

Rosenbaum, P. R. (2010). *Design of observational studies*. New York: Springer.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*, 41-55. doi:10.2307/2335942

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician, 39*(1), 33-38. doi:10.1080/00031305.1985.10479383

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

Rossell, C. H., & Kuder, J. (2005). Mega-murky: A rebuttal to recent meta-analyses on bilingual education. In J. Sohn (Ed.), *The effectiveness of bilingual school programs for immigrant children* (pp. 43-76). Berlin.

Rubin, D. B. (1973). Matching to remove bias in observational studies. *Biometrics, 29*, 159-183. doi:10.2307/2529684

Rubin, D. B. (2001). Using propensity scores to help design observational studies: Application to the tobacco litigation. *Health Services and Outcomes Research Methodology, 2*(3), 169-188. doi:10.1023/A:1020363010465

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association, 100*(469), 322-331. doi:10.1198/016214504000001880

Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine, 26*(1), 20-36. doi:10.1002/sim.2739

Rubin, D. B., & Thomas, N. (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics, 52*(1), 249-264. doi:10.2307/2533160

Ruiz Soto, A. G., Hooker, S., & Hatalova, J. (2015). *Top languages spoken by English language learners nationally and by state*. Retrieved from http://www.migrationpolicy.org/research/top-languages-spoken-english-language-learners-nationally-and-state

Saiz, A., & Zoido, E. (2005). Listening to what the world says: Bilingualism and earnings in the United States. *Review of Economics and Statistics, 87*, 523-538. doi:10.1162/0034653054638256

Saunders, W. M., & O'Brien, G. (2006). Oral language. In F. Genesee, K. Lindholm-Leary, W. M. Saunders, & D. Christian (Eds.), *Educating English language learners*. New York: Cambridge University Press.

Schatschneider, C., Fletcher, J. M., Francis, D. J., Carlson, C. D., & Foorman, B. R. (2004). Kindergarten prediction of reading skills: A longitudinal comparative analysis. *Journal of Educational Psychology, 96*(2), 265-282. doi:10.1037/0022-0663.96.2.265

Schlesinger, A. M. (1991). *The disuniting of America: Reflections on a multicultural society*. New York: W. W. Norton & Company.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference (pp. 33-63). Belmont, CA: Wadsworth Cengage Learning.

Shannon, S. M., & Milian, M. (2002). Parents choose dual language programs in Colorado: A survey. *Bilingual Research Journal, 26*(3), 681-696. doi:10.1080/15235882.2002.10162584

Siegel, L. S., & Ryan, E. B. (1989). The development of working memory in normally achieving and subtypes of learning disabled children. *Child Development, 60*(4), 973-980. doi:10.2307/1131037

Silver, L. B. (1989). Psychological and family problems associated with learning disabilities: Assessment and intervention. *Journal of the American Academy of Child & Adolescent Psychiatry, 28*(3), 319-325. doi:10.1097/00004583-198905000-00003

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research, 75*(2), 247-284. doi:10.3102/00346543075002247

Slavin, R. E., Madden, N., Calderón, M., Chamberlain, A., & Hennessy, M. (2011). Reading and language outcomes of a multiyear randomized evaluation of transitional bilingual education. *Educational Evaluation and Policy Analysis, 33*(1), 47-58. doi:10.3102/0162373711398127

Snow, C. (2006). Cross-cutting themes and future research directions. In D. August & T. Shanahan (Eds.), *Developing literacy in second-language learners: Report of the National Literacy Panel on Language-Minority Children and Youth.* Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Snyder, T. D., De Brey, C., & Dillow, S. A. (2019). *Digest of Education Statistics 2017*. Retrieved from https://nces.ed.gov/pubs2018/2018070.pdf

Speich, B., von Niederhäusern, B., Schur, N., Hemkens, L. G., Fürst, T., Bhatnagar, N., . . . Briel, M. (2018). Systematic review on costs and resource use of randomized clinical trials shows a lack of transparent and comprehensive data. *Journal of Clinical Epidemiology, 96*, 1-11. doi:10.1016/j.jclinepi.2017.12.018

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science, 25*, 1-21. doi:10.1214/09-STS313

Sullivan, G. M. (2011). Getting off the "gold standard": Randomized controlled trials and education research. *Journal of Graduate Medical Education, 3*(3), 285-289. doi:10.4300/JGME-D-11-00147.1

Swanson, H. L., Sáez, L., & Gerber, M. (2006). Growth in literacy and cognition in bilingual children at risk or not at risk for reading disabilities. *Journal of Educational Psychology, 98*(2), 247-264. doi:10.1037/0022-0663.98.2.247

Thoemmes, F. (2012). Propensity score matching in SPSS. Retrieved from https://arxiv.org/abs/1201.6385

Thoemmes, F., & Kim, E. S. (2011). A systematic review of propensity score methods in the social sciences. *Multivariate Behavioral Research, 46*(1), 90-118. doi:10.1080/00273171.2011.540475

Thomas, W. P., & Collier, V. P. (1997a). *School effectiveness for language minority students*. Retrieved from NCBE website: www.ncbe.gwu.edu

Thomas, W. P., & Collier, V. P. (1997b). Two languages are better than one. *Educational Leadership, 55*(4), 23.

Torgesen, J. K., & Bryant, B. R. (1995). *Test of Phonological Awareness*. Austin, TX: PRO-ED.

Townsend, D., & Collins, P. (2008). English or Spanish? Assessing Latino/a children in the home and school languages for risk of reading disabilities. *Topics in Language Disorders, 28*(1), 61. doi:10.1097/01.adt.0000311416.77590.b5

Relating to bilingual education and English as a second language and other special language programs in public schools, Texas State Senate Bill 477 (1981).

U.S. Department of Education. (2003). *Identifying and implementing educational practices supported by rigorous evidence: A user friendly guide*. Retrieved from https://www2.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf

U.S. Department of Education. (2010). *A blueprint for reform: The reauthorization of the Elementary and Secondary Educaton Act*. Retrieved from https://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

Ulloa, J. (2016). California will bring back bilingual education as Proposition 58 cruises to victory. *The Los Angeles Times*. Retrieved from http://www.latimes.com/nation/politics/trailguide/la-na-election-day-2016-proposition-58-bilingual-1478220414-htmlstory.html

United States Census Bureau. (2018). *Historical poverty tables: People and families - 1959-2017*. Retrieved from: https://www.census.gov/data/tables/time-series/demo/income-poverty/historical-poverty-people.html

United States Department of Commerce. (2012). Annual Social and Economic (ASEC) Supplement to the Current Population Survey (CPS). http://www.census.gov/population/hispanic/data/2012.html

Unz, R., & Ruchman, G. (1997). *English language education for children in public schools*. Retrieved from http://www.languagepolicy.net/archives/unztext.htm

Vaughn, S., Cirino, P. T., Linan-Thompson, S., Mathes, P. G., Carlson, C. D., Hagan, E. C., . . . Francis, D. J. (2006). Effectiveness of a Spanish intervention and an English intervention for English-language learners at risk for reading problems. *American Educational Research Journal, 43*(3), 449-479. doi.org/10.3102/00028312043003449

Vaughn, S., Cirino, P. T., Tolar, T., Fletcher, J. M., Cardenas-Hagan, E., Carlson, C. D., & Francis, D. J. (2008). Long-term follow-up of Spanish and English interventions for first-grade English language learners at risk for reading problems. *Journal of Research on Educational Effectiveness, 1*(3), 179-214. doi:10.1080/19345740802114749

Wagner, R. K., Torgensen, J., & Rashotte, C. A. (1999). *Comprehensive Test of Phonological Processing*. Austin, TX: Pro-Ed.

Waitoller, F. R., Artiles, A. J., & Cheney, D. A. (2010). The miner's canary: A review of overrepresentation research and explanations. *The Journal of Special Education, 44*(1), 29-49. doi:10.1177/0022466908329226

Wang, S. V., Jin, Y., Fireman, B., Gruber, S., He, M., Wyss, R., . . . Gagne, J. J. (2018). Relative performance of propensity score matching strategies for subgroup analyses. *American Journal of Epidemiology, 187*(8), 1799-1807. doi:10.1093/aje/kwy049

Wiese, A.-M., & García, E. E. (1998). The Bilingual Education Act: Language minority students and equal educational opportunity. *Bilingual Research Journal, 22*(1), 1-18. doi:10.1080/15235882.1998.10668670

Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery - Revised*. Itasca, IL: Riverside.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1995). *Woodcock Language Proficiency Battery Revised, Spanish Form: Supplemental Manual*. Itasca, IL: Riverside.

Yow, W. Q., & Li, X. (2015). Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: Why variations in bilingual experiences matter. *Frontiers in Psychology, 6*, 164. doi:10.3389/fpsyg.2015.00164

Ziegler, J. C., & Goswami, U. (2005). Reading acquisition, developmental dyslexia, and skilled reading across languages: A psycholinguistic grain size theory. *Psychological Bulletin, 131*(1), 3-29. doi:10.1037/0033-2909.131.1.3

Appendix

Definition of Terms

Bilingual education. This instructional approach for ELs is designed to develop academic skills in both the native language and English. There are three common models of bilingual education: early transition bilingual, late transition bilingual, and dual language (Ochoa, 2005).

Dual language programming. Also known as "two-way," this instructional program is offered to equal numbers of both English-speaking and EL students who seek to become bilingual over the course of at least the first four-to-six years of schooling in this model. Among a variety of dual language models is that in which the language of instruction varies by content area between the ELs' home language and English, achieving a 50-50 balance. Despite great variation in implementation, the common element is that reading instruction is provided in both languages for all students (Goldenberg & Coleman, 2010).

Early transition bilingual programming. Also known as "early-exit" or "transitional," this instructional program is typically offered for the first two-to-four years of school to a classroom of EL students. Spanish reading instruction is provided early on as a bridge to English reading, but Spanish may not appear in other content areas outside of reading (Slavin et al., 2011).

English Learner (EL). A student who is acquiring English as a second language (LaCelle-Peterson & Rivera, 1994).

Immersion education. This instructional approach for ELs is designed to develop academic skills in English only. Native language skills are not targeted. There are two categories of common models of immersion education: traditional immersion and pull-out immersion (Ochoa, 2005).

Late transition bilingual programming. Also known as "maintenance," "late-exit," or "developmental," this instructional program is typically offered for the first four-to-six years of school to a classroom of EL students. Instruction in the early years is delivered in Spanish with some English language instruction and, over time, the balance gradually shifts to primarily English (Genesee et al., 2004).

Limited English Proficient (LEP). A federal category of persons entitled to language assistance in accessing a service or benefit that is federally-funded, such as public education, per the Title VI regulations of the Civil Rights Act of 1964 (Lhamon & Gupta, 2015, January 7)

Oral language. Oral language proficiency is defined as both receptive and expressive, encompassing knowledge or use of specific components, such as phonology, vocabulary, grammar, and pragmatic skills (Lesaux & Geva, 2006).

Pull-out immersion programming. Explicit English language instruction, as opposed to content-based, is offered outside of the classroom and is supplemental to the child's regular instruction. This program may be available only to "newcomers" to the country, regardless of age, to ease their transition to American schooling, and typically excludes American children from Spanish-speaking homes (Wiese & García, 1998).

Phonological awareness (PA). This is the knowledge that words are composed of sounds and has been defined as "the ability to consciously attend to the sounds of language as distinct from its meaning" (Lesaux & Geva, 2006, p. 55).

Traditional immersion programming. Also known as "content-based" or "sheltered English," this instructional program focuses on delivering academic content in English but in an accessible manner, using gestures and other visual cues to assist. It is typically

offered as needed, regardless of age. Models vary from students spending half to entire days in this setting, and the goal is English acquisition; native language maintenance is not facilitated (Wiese & García, 1998).