ELSEVIER

Early Childhood Research Quarterly 22 (2007) 277–293

Early Childhood Research Quarterly

Two-way and monolingual English immersion in preschool education: An experimental comparison

W. Steven Barnett*, Donald J. Yarosz, Jessica Thomas, Kwanghee Jung, Dulce Blanco

National Institute for Early Education Research, Rutgers, The State University of New Jersey, United States

Received 20 June 2006; received in revised form 15 March 2007; accepted 23 March 2007

Abstract

An experimental study was conducted comparing the effects of dual language, or two-way immersion (TWI) and monolingual English immersion (EI) preschool education programs on children's learning. Three-and four-year old children were randomly assigned by lottery to either a newly established TWI Spanish/English program or a monolingual English program in the same district. Children in the study were from both Spanish and English home language backgrounds. All classrooms in the study used the High/Scope curriculum, and all met high standards for teacher qualifications, ratio, and class size. The TWI program alternated between English and Spanish weekly by rotating children between two classrooms (and teachers) each week. Programs were compared on measures of children's growth in language, emergent literacy, and mathematics. Children in both types of classrooms experienced substantial gains in language, literacy, and mathematics. No significant differences between treatment groups were found on English language measures. Among the native Spanish speakers, the TWI program produced large gains in Spanish vocabulary compared to the EI program. Both TWI and EI approaches boosted the learning and development of children including ELL students, as judged by standard score gains. TWI also improved the Spanish language development of English language learners (ELL) and native English speaking children without losses in English language learning.
© 2007 Elsevier Inc. All rights reserved.

Keywords: Preschool education; Bilingual education; Early childhood policy; Randomized trial

1. Introduction

The number of children in the United States who are English language learners (ELL) is growing rapidly, and the vast majority are from homes where Spanish is the primary language (Tabors, Paez, & Lopez, 2003). This trend is even stronger at the preschool level. Hispanic children account for more than 20% of all children under five (Collins & Ribeiro, 2004). Moreover, Hispanic children are more highly represented in public preschool programs because they are more likely to be income eligible for such programs. About 30% of the Head Start population is Hispanic (Collins & Ribeiro, 2004). In the public schools, about 25% of preschool children are Hispanic (Smith, Kleiner, Parsad, & Farris, 2003). Head Start data indicate that Spanish is the dominant language for three-quarters of these children (Collins & Ribeiro, 2004). A number of studies have found preschool education to improve the school readiness of

* Corresponding author at: NIEER, 120 Albany Street, Suite 500 New Brunswick, NJ 08901, United States. Tel.: +1 732 932 4350; fax: +1 732 932 4360.

E-mail address: sbarnett@nieer.org (W.S. Barnett).

0885-2006/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ecresq.2007.03.003

Hispanic children who experience high rates of school failure and high school dropout (Gormley, Gayer, Phillips, & Dawson, 2005; Johnson & Walker, 1991; Puma et al., 2005). These children present challenges and opportunities for public preschool education programs (Barnett, Hustedt, Robin, & Schulman, 2005).

A key issue in the education of ELL children is language of instruction. School systems historically have encouraged English language abilities while neglecting a child's home language (Hakuta, 1986). However, a variety of bilingual approaches have been developed with the goal of increasing the achievement of ELL children. Research tends to find that bilingual and first language education are at least as effective as English immersion (EI), but these approaches remain controversial (Greene, 1997; Rossell & Baker, 1996; Slavin & Cheung, 2005; Willig, 1985).

Some have suggested that children may lose their first language when a "high-prestige" second language is introduced, arguing for emphasis on the first language in preschool education (Wong Filmore, 1991). Yet, studies have found that bilingual preschool programs promoted development in both languages rather than impeding growth in the first language (Winsler, Diaz, Espinosa, & Rodriquez, 1999). The outcomes appear to depend on how well development in each language is supported (Tabors & Snow, 2001).

A recent meta-analysis of 17 studies of K-12 program effectiveness with ELL students published since Willig's (1985) early meta-analysis found no advantage for all-English instruction (Rolstad, Mahoney, & Glass, 2005). The average effect sizes (es) for bilingual approaches compared to others was .08, but bilingual programs that supported academic development in both languages (es = 23) outperformed bilingual programs designed to transition students to English (es = −.01). Studies had larger effects if they compared only ELL students (es = .23) than if they compared ELL students to language majority students (es = .05). The effect size was much larger, .86, for outcomes measured in the native language.

Two-way bilingual immersion (TWI) is a promising and increasingly common approach to addressing the needs of ELL Students (Howard, Sugarman, & Christian, 2003; Lindholm-Leary, 2005; Short, 1993). TWI (also called dual language) programs provide ELL and native English speakers with an education in two languages. Immersion is a particularly powerful approach for young children whose success depends on large amounts of exposure to oral language (DeKeyser & Larson-Hall, 2005; Johnson & Newport, 1989; Lanauze & Snow, 1989). Such programs build on research on effective practices in the education of ELL students, including evidence that first language proficiency strengthens second language acquisition and that bilingualism has cognitive benefits (August & Hakuta, 1997; Cummins, 1981; Reynolds, 1991). Although much of this research has been conducted with older children, there is similar evidence from the preschool years (Campbell & Sais, 1995; Campos, 1991; Diaz, Padilla, & Weathersby, 1991). Children who become bilingual can later surpass monolingual children in cognitive processing skills, and early oral language facility in the first language can facilitate later reading proficiency in English (Bialystok, 2006; Bialystok & Martin, 2004). In addition, bilingualism is valuable *per se* given the importance of global interactions in culture, politics, and economics, and TWI programs seek to build on the strengths of ELL students and redress the problem of America's traditionally weak foreign language programs by increasing bilingual proficiency for children regardless of their primary language (Howard et al., 2003).

Nevertheless, additional rigorous research is needed to more fully specify the most effective approaches to preschool education for ELL children, including the rapidly growing Spanish-language population (Bialystok & Herman, 1999; Dickinson & Tabors, 2001; Garcia, 2000; Tabors & Snow, 2001). Past studies with preschool children have suggested that bilingual, and TWI approaches, in particular, may have important advantages (e.g., Rodriguez, Diaz, Duran, & Espinosa, 1995; Stipek, Ryan, & Alarcon, 2001; Tabors et al., 2003; Winsler et al., 1999). As noted earlier, these benefits include the extensive exposure to language such approaches provide which is so essential for young children's language learning, the greater language proficiency acquired due to such early exposure, the benefits of bilingualism for cognitive abilities, and long-term cross over benefits for English language proficiency (see also, Reese, Garnier, Gallimore, & Goldenberg, 2000). To our knowledge, our study is the first to use random assignment of both ELL and other children to either a TWI or an English immersion preschool program to compare the educational effectiveness of these two approaches.

2. Method

The study was conducted in a Northeastern city with a population of just over 120,000 in which 50% of the population was Latino (U.S. Census Bureau, 2000). Median household income was $35,175 in 1999, and 21% of the children under age 5 were in poverty. In 2004–2005, 76% of public school children qualified for free or reduced price lunch. One

school in the city was selected to implement a new TWI preschool program. Otherwise children attended programs that were primarily in English (EI), though some support was provided in these programs for home language development in Spanish and other languages. A lottery for admission to the TWI program that was held because citywide demand for the program greatly exceeded the supply of spaces created the opportunity for randomization.

2.1. Sample and program assignment

In the fall of 2002, nearly 1000 3- and 4-year olds applied to the TWI program. All applicants were entered into a lottery to select 150 3-year olds and 150 4-year olds for attendance. Each applicant had an equal chance (within the two age groups) to attend. We randomly selected 50 of the lottery "winners" at each age for the TWI study group. Of these, 85 agreed to participate in the study (i.e., permit their children to be tested, etc.). From the list of lottery entrants not assigned to the TWI program and locatable, we randomly selected 127 for invitation to the study. Of these, 62 families agreed to participate. The lower percentage of participation by control (EI) families may be due to disappointment at not winning admission to the TWI program. The number of children tested at post-test as well as pre-test was somewhat smaller (79 TWI and 52 EI) primarily because children's families moved away (one EI child declined testing).

2.2. Program description

In the TWI program, each classroom of students rotates weekly between two teaching teams. One team consists of a teacher and an assistant who teach in English. The other team consists of a teacher and an assistant who teach in Spanish. Thus, each week half of the students in the school are taught in English and half are taught in Spanish. Members of both teaching teams collaborate to plan curricula and share assessment data for the children they jointly teach. Students who speak primarily English are integrated with students whose first language is not English. In our sample, 57% of children spoke primarily Spanish at home, 37% spoke primarily English, and 5% had another primary language. Study children were enrolled in 20 (10 pairs of English and Spanish) TWI classrooms and 16 EI classrooms (with some bilingual support services).

All classrooms in the study employed a licensed teacher and an assistant with a Child Development Associate (CDA) credential to work with a maximum of 15 children per classroom. All classrooms used the High Scope curriculum (Hohmann & Weikart, 1995), which has been adopted citywide by the public schools. The High/Scope preschool curriculum is an "open framework" approach that is widely used in the United States. The curriculum has a Piagetian theoretical foundation and emphasizes active learning in both teacher-and child-initiated activities in 10 content areas within which there are key experiences. A plan-do-review sequence for children's play is a central element of the daily routine. Several studies have found the High/Scope curriculum effective in improving the learning and development of disadvantaged children (Frede & Barnett, 1992; Schweinhart et al., 2005; Smith, 1975).

The TWI programs operated for 8 h a day over 200 school days. The comparison classrooms operated for 6 h a day over 180 school days. Both programs had wrap-around child care to extend the day up to 10 h and to cover up to 245 days. For comparison purposes, data were collected on teachers and on classroom quality for all 20 TWI classrooms and all 16 EI classrooms. Teachers did not significantly differ in experience, field of certification, or whether they took the traditional or alternate route to teacher certification. However, TWI teachers were more likely to have a Masters degree (30% versus 6%) and to be bilingual (50% versus 25%).

2.3. Child assessment

Children were tested in the fall and spring of the 2003–2004 school year. Fall testing took place over 5 weeks, and spring testing spanned 4 weeks. The TWI group was administered each measure in both English and Spanish. In the EI group, Spanish-dominant children were assessed in English and Spanish; the others were tested in English only. Children tested in both languages were tested first in their home language. Primary language was ascertained from the classroom teacher who made judgments based on language proficiency tests administered by the schools at the beginning of the year, supplemented by parental report and their own experiences with the children. Assessors spoke to the child only in the language of each assessment to avoid code switching during testing sessions. Assessments were conducted one-on-one in the child's school, and assessments were scheduled to avoid meal, nap and outdoor play times. Testing sessions lasted 20–30 min.

For assessment, children who had a primary language other than English or Spanish were combined with the native English speakers. Ideally, these children would have constituted a separate group for assessment and analysis. However, given the small number this was not practical, and the alternative would have been to delete these children from the study. We emphasize that these children were found to have sufficient proficiency in English to participate in the English language assessments and did not significantly differ from children whose primary language was English on the pre-test measures of language and literacy.

There were two teams of assessors, one for each language. They included research associates with years of experience testing children and less experienced recent college graduates and advanced teacher education students. Assessors were trained on each assessment and then shadow scored in practice assessments until they reached 100% agreement with the trainer, who also was the site coordinator. The site coordinator was responsible for assuring adequate reliability throughout the study.

Individualized assessments were selected to measure the contributions of the preschool programs to children's learning, with emphasis on language and emergent literacy skills. Criteria for instrument selection included: (1) availability of equivalent tasks in both Spanish and English; (2) reliability and validity, particularly pre-literacy skills that are good predictors of later reading ability; and (3) appropriateness for children ages 3–5. Each assessment instrument is discussed below.

2.4. The Peabody picture vocabulary test (PPVT–III)

The PPVT–III (Dunn & Dunn, 1997) is a 204-item test of receptive vocabulary in standard English. The test is administered by having children point to one of four pictures shown when given a word to identify. The PPVT-III is often used as a quick indicator of general cognitive ability, and it correlates reasonably well with other measures of linguistic and cognitive development related to school success. The PPVT-III has a mean standard score of 100 and a standard deviation of 15. The PPVT-III was nationally standardized on a stratified normative sample of 2000 children and adolescents and has an internal consistency reported as Spearman–Brown split half reliability coefficients ranging from .92 to .98. Test-retest reliability for a 1-month interval in four different age groups ranged from .91 to .93.

2.5. Test de Vocabulario en Imagenes Peabody (TVIP)

The TVIP (Dunn, Lugo, Padilla, & Dunn, 1986) uses 125 translated items from the PPVT to assess receptive vocabulary acquisition of Spanish-speaking and bilingual students. Items were selected through item analysis for their universality and appropriateness to Spanish-speaking communities, according to publishers. The test is appropriate for measuring growth in Spanish vocabulary for bilingual students and for those who do not come from Spanish language homes, according to its authors. The TVIP has a mean standard score of 100 and a standard deviation of 15. The TVIP was standardized on two monolingual, Spanish-speaking populations (1,219 children from public schools in Mexico and 1,488 children from the San Juan metropolitan area of Puerto Rico). Internal consistency reliability for this measure is reported to be .91 to .94 in the relevant age range. Correlations between the TVIP and Spanish language tests of general ability range from .25 to .56.

2.6. The Woodcock-Johnson psycho-educational battery-revised (WJ-R) and the Bateria Psico-Educativa Revisada de Woodcock-Muñoz-Revisada (WM-R)

The WJ-R and WM-R (Woodcock & Johnson, 1989; Woodcock & Munoz-Sandoval, 1996) are comprehensives sets of individually administered tests of cognitive abilities and achievement. We administered the Picture Vocabulary and Applied Problems subtests from these batteries in both English and Spanish. The 58-item Picture Vocabulary (Vocabulario con Dibujos) subtest asks children to select pictures to match words and to say a word when shown a picture. It is primarily an expressive vocabulary task that measures the ability to name familiar and unfamiliar pictured objects. The 60-item Applied Problems (Problemas Aplicados) subtest measures math skills.

The WJ-R and WM-R tests were calibrated and equated to U.S. norms through Rasch modeling and are particularly well-suited to the needs of assessment with bilingual populations (Woodcock & Munoz-Sandoval, 1996). The tests' standard scores have a mean of 100 and a standard deviation of 15. The English Form of the subtests was normed on a stratified random sample of 6,359 English-speaking subjects in the United States. The Spanish Form was normed

on 3,911 primarily monolingual Spanish-speaking subjects from samples obtained both inside and outside the United States. Internal consistency reliabilities range from the high .70 s to low .90 s on both subtests for preschool-aged children. Correlations of the WJ-R and WM-R with other tests of cognitive ability and achievement are reported to range from .60 to .70.

2.7. Measures of emergent literacy skills

Few measures of literacy skills for preschool children are available in English and Spanish and there are questions about the comparability of English and Spanish language versions of such tests and how well they measure the relevant skills of young children (Huempfner, 2004; Lopez & Greenfield, 2004). These concerns led us to employ some new measures that we hoped might provide better assessments of children's literacy skills. We administered two phonological awareness (phoneme deletion and rhyme recognition) subtests in both English and Spanish. These criterion-referenced tests are early versions of tests developed specifically to fill the need for equivalent tests in English and Spanish that could be used to study the literacy development of bilingual young children (Lopez, 2005; Lopez & Greenfield, 2004). Evidence of concurrent validity is provided by correlations of .33 to .52 with the Spanish and English PreLas2000 (Lopez, 2005; Lopez & Greenfield, 2004). The tests measure skills that are highly predictive of later reading abilities and are important milestones in the literacy development of preschool children (Snow, Burns, & Griffin, 1998). Phonological awareness and letter recognition are believed to play key roles in children becoming successful readers (Scarborough, 1998; Whitehurst & Lonigan, 2001). Our study found some evidence of reasonable psychometric properties, but also some limitations in these new measures. In general, they performed more poorly than the older, standardized assessments.

The *Phoneme Deletion* test consists of three practice items followed by eight test items. A puppet is introduced and the child is shown a target picture and given its name (e.g., this is a rock). The child is asked to repeat the name of the picture and then the puppet repeats the name (sometimes correctly, sometimes deleting beginning or ending sounds). The child has to judge if the puppet is correct or not and then correct the puppet, if necessary. The child must pass at least one of the three practice items for the subtest to be administered. This test assesses a child's ability to discern correct versus incorrect pronunciation of basic vocabulary in both English and Spanish. Correlations from pre-to post-test were .37 and .32 for the English and Spanish versions, respectively. The Spanish and English language scores on this test were more moderately correlated with each other, .49 at pre-test and .32 at post-test. Correlations with age were .51 and .47 for the English version at pre-and post-test and .53 for the Spanish version at both times, indicating that scores increased significantly with age. The Spanish version correlated .30 at pre-test and .47 at post-test with the TVIP, and the English version correlated .55 at pre-test and .48 at post-test with the PPVT. All correlations were significant at $p < .05$. Scatter plots indicated a ceiling problem (topping out on the test) at post-test.

The *Rhyme Recognition* subtest consists of three practice items followed by eight test items. Children are shown three pictures on a page and are asked to point to the two that rhyme after repeating all of them first with the examiner. The child needs to pass at least one of the practice items in order for the subtest to be administered. Of all the instruments we used, this one was most problematic due to its difficulty. The rhyme recognition tests proved too difficult for most children at pre-test, and remained difficult for many children at post-test when 51% could not complete the task in Spanish and 65% could not complete the task in English. By contrast, at post-test in either language the phoneme deletion task proved too difficult for only 5% of the children, and the letter recognition tasks were too difficult for only about 10% of the children. Given the high percentage of floor scores, analyses were conducted with dummy variables entered to indicate floor scores, but these did not yield appreciably different results and the simpler regressions without these are reported.

For the Rhyme measure, correlations (all significant at $p < .05$) from pre-to post-test were .56 and .59 for the English and Spanish versions, respectively. The Spanish and English language scores on this test were fairly highly correlated with each other, .76 at pre-test and .59 at post-test. Correlations with age were .37 and .45 for the English version and .40 and .35 for the Spanish version at pre-and post-test, respectively. The Spanish version correlated .25 at pre-test and .32 at post-test with the TVIP, and the English version correlated .52 at pre-test and .48 at post-test with the PPVT.

An *Alphabet Recognition* test was developed at the National Institute for Early Education Research (NIEER) to assess letter recognition in English and Spanish (Blanco, 2004). This measure allows children to identify all of the letters that they might know; other letter recognition tests commonly present only a small subset of the alphabet to

children. This assessment is similar to the Head Start National Reporting System Letter Naming measure, which is reported to have IRT reliability coefficients of .90 in English and .80 in Spanish (Head Start National Reporting System, 2006). A child is presented two letters of the alphabet on each page and is asked to name one of them at a time. The answer is considered correct if the child gives a correct letter name response, a correct letter sound response or a word starting with the letter presented. The score is calculated as the total number of correct responses. Pre-to post-test correlations were .71 for the English and .93 for the Spanish versions. The Spanish and English language test scores correlated with each other were .78 at pre-test and .93 at post-test.

Looking at indications of validity, correlations with age were .37 and .29 for the English version and .27 and .33 for the Spanish version at pre-and post-test, respectively. Correlations of the Spanish version with the TVIP were not statistically significant, while correlations of the English version with the PPVT were .51 and .41 at pre-and post-test. Interestingly, the Spanish letter naming task scores also were significantly correlated with the PPVT scores, .44 and .47 at pre-and post-test, respectively.

2.8. Measurement of classroom environment and activities

To identify similarities and differences between the two types of preschool programs, we conducted classroom observations using three rating scales. Observers generally were students with a background in education or psychology and former preschool teachers. Observers were trained on the classroom measures and then shadow scored on practice observations until they reached an 85% inter-rater reliability on each instrument with the trainer. They were shadow scored again half way through data collection (observers worked on more than one study during the spring) to check for drift, and reliability in these checks averaged about .90. All observations were conducted during February and March. All TWI classrooms ($N = 20$) and EI classrooms ($N = 16$) were observed once in the morning for about 4 h. Each observation period included greeting and one meal. Observers scored all of the instruments immediately after completing the observation.

2.9. Early childhood environmental rating scale-revised (ECERS-R)

The ECERS-R (Harms, Clifford, & Cryer, 1998) provides a global measure of preschool classroom quality with 43 items that cover a broad range of quality considerations from safety to teacher–child interaction to parent involvement. The ECERS-R scale uses a scoring system where 1 is considered inadequate, 3 minimal, 5 good, and 7 excellent. The ECERS-R (and its predecessor, the ECERS) has been widely used to measure quality in research studies within the United States and abroad. Scores have been found to correlate moderately with more specific measures of teaching quality and to predict children's learning and development (Peisner-Feinberg & Burchinal, 1997; Sylva & Siraj-Blatchford, 2001). The total scale internal consistency for the ECERS-R is .92, and subscale internal consistencies range from .71 to .88.

2.10. Supports for early literacy assessment (SELA)

The SELA (Smith, Davidson, Weisenfeld, & Katsaros, 2001) was used to measure the quality of the preschool literacy environment and instruction in greater detail. The SELA consists of 16 items relating to the literacy environment, language development, print/book concepts, phonological awareness, letters and words, and parent involvement. Scores on the SELA range from 1 to 5, with 1 considered very low quality, 3 fair quality, and 5 ideal quality. The SELA was developed based on current research and professional opinion regarding best practices with regard to literacy. We employed the SELA because of concerns that the ECERS-R devotes insufficient attention to the "academic" aspects of preschool education, including literacy development (Sylva et al., 2006). We found adequate reliability of the SELA, as indicated by a Cronbach's alpha of .84. In another study, with 310 classrooms, the correlation between the SELA and ECERS-R total scores was found to be .75, which provides some validation of the SELA while suggesting that it measures somewhat different aspects of the classroom environment (Lamy, 2004). This result is virtually identical to the correlation between the ECERS-R and the "ECERS-Extension," which was developed to focus added attention on literacy, math, science, and diversity (Sylva et al., 2006). The SELA also has been shown to be sensitive to changes in classroom practices over time (Lamy et al., 2005).

Table 1
Sample description

	Two-way immersion		English immersion	
	Number	Percent	Number	Percent
Gender				
Male	31	39.2	26	50.0
Female	48	60.8	26	50.0
Child's age				
3 years old	38	48.1	18	34.6
4 years old	41	51.9	34	65.4
Home language				
Spanish	47	59.5	28	53.8
English	28	35.4	21	40.4
Other	4	5.1	3	5.8
Race-ethnicity				
White/non-Hispanic	6	7.6	4	7.7
African American	9	11.4	8	15.4
Hispanic	62	78.5	38	73.1
Other	2	2.5	1	1.9

2.11. Supports for English language learners classroom assessment (SELLCA)

The SELLCA (NIEER, 2005) was developed to assess the extent to which the preschool teacher and classroom environment support the first language (L1) development of ELL children. Scores on SELLCA items range from 1 to 5, with 1 considered very low, 3 fair, and 5 ideal. We used three items from the SELLCA to measure teacher practices in all classrooms as indicators of the extent to which actual treatment corresponded to intent and to compare the EI, English language TWI, and Spanish language TWI classrooms. These three indicators provided ratings for: teachers' use of children's primary language; degree to which children's cultural backgrounds are incorporated in the classroom; and, use of effective strategies to support English language development.

3. Results

Descriptive statistics for the analyzed sample are displayed in Table 1. Pre-test and post-test scores for all child outcome measures are reported in Tables 2 and 3. A small number ($n = 7$) of children had a primary language other than English or Spanish. It is noteworthy that their parents applied for them to enter the dual-language English–Spanish preschool program (the only way they would appear in the study), and all of them had sufficient facility in English to be tested on the English language measures. They did not significantly differ from those students who had English as a primary language on any of the English language measures at pre-test.

The TWI and EI groups did not significantly differ in their demographic characteristics (independent samples *t*-tests) or on any of the pre-tests (as determined by ANOVA). Nevertheless, analyses of the post-test scores presented later control for age, language, gender and pre-test to increase statistical power and because modest (though not significant) differences between TWI and EI groups on these variables might influence results of the analyses.

3.1. Classroom quality

As reported in Table 4, the average score was "good" on the ECERS-R and "fair" on the SELA. The ECERS-R scores are better than has been found in larger studies of preschool programs across the United States (National Center for Early Development and Learning (NCEDL), 2005). *t*-Tests revealed no significant differences in mean ECERS-R or SELA scores between TWI and EI classrooms. ANOVA was used to compare TWI English, TWI Spanish, and EI classrooms on the three indicators taken from the SELLCA. The ANOVA found significant differences between the three types of classrooms for teachers use of the children's primary language ($F(2,33) = 34.87$, $p < .05$) and support for children's cultural background ($F(2,33) = 10.75$, $p < .05$). Post hoc Scheffe tests were conducted to identify which types

Table 2
Pre- and post-test child outcome measures—English measures, full sample

	English immersion			Two-way immersion		
	Pre-test *M* (S.D.)	Post-test *M* (S.D.)	*N*	Pre-test *M* (S.D.)	Post-test *M* (S.D.)	*N*
PPVT raw score	30.5 (17.3)	41.3 (16.7)	52	29.3 (16.5)	41.1 (16.5)	79
PPVT standard score	77.9 (17.2)	83.7 (13.3)	52	78.2 (16.0)	84.2 (12.4)	79
WJ-R PV raw score	15.7 (5.3)	18.9 (4.2)	52	15.5 (5.6)	18.6 (3.4)	79
WJ-R PV standard score	78.6 (18.1)	85.1 (17.6)	52	80.1 (18.0)	83.6 (14.2)	79
WJ-R AP raw score	6.8 (3.9)	9.6 (4.1)	49	6.2 (4.4)	10.3 (6.7)	79
WJ-R AP standard score	90.6 (12.0)	92.5 (12.1)	49	89.2 (13.7)	93.1 (11.6)	79
PD-E	9.0 (4.4)	11.1 (4.1)	51	8.8 (4.7)	11.8 (3.1)	79
Letter ID-E	4.8 (7.6)	11.0 (9.1)	52	4.1 (6.1)	9.9 (8.9)	79
Rhyme-E	0.8 (1.7)	1.6 (2.3)	52	1.0 (2.0)	1.5 (2.5)	79

PPVT—the Peabody picture vocabulary test.
WJ-R PV—Woodcock Johnson picture vocabulary.
WJ-R AP—Woodcock Johnson applied problems
PD-E—Phoneme deletion-English.
Letter ID-E—alphabet recognition-English.
Rhyme-E—rhyme recognition-English.

Table 3
Pre-and Post-test child outcome measures—Spanish measures, Spanish speaking only

	English immersion			Two-way immersion		
	Pre-test *M* (S.D.)	Post-test *M* (S.D.)	*N*	Pre-test *M* (S.D.)	Post-test *M* (S.D.)	*N*
TVIP raw	12.0 (11.2)	14.9 (10.7)	27	12.1 (9.0)	21.3 (12.2)	47
TVIP standard	83.4 (12.4)	79.6 (12.4)	27	83.8 (13.1)	87.4 (15.7)	47
WM-R PV raw	13.3 (3.9)	12.7 (4.0)	27	14.3 (3.2)	13.7 (3.6)	47
WM-R PV standard	73.3 (12.0)	63.4 (13.4)	27	74.7 (14.8)	65.5 (16.7)	47
WM-R AP raw	5.0 (3.2)	8.3 (3.8)	27	5.5 (3.0)	8.7 (2.5)	47
WM-R AP standard	84.2 (9.7)	88.2 (10.4)	27	86.1 (10.1)	87.7 (9.7)	47
PD-S	9.0 (4.7)	11.4 (3.0)	27	8.9 (4.4)	12.3 (3.0)	47
Letter ID-S	2.7 (5.6)	10.3 (9.4)	27	3.3 (5.2)	7.9 (7.2)	47
Rhyme-S	0.9 (1.8)	1.5 (2.4)	27	0.6 (2.0)	2.6 (2.8)	47

TVIP—Test de Vocabulario en Imagenes Peabody.
WM-R PV—Woodcock-Muñoz picture vocabulary.
WM-R AP—Woodcock-Muñoz applied problems.
PD-S—Phoneme deletion-Spanish.
Letter ID-S—alphabet recognition-Spanish.
Rhyme-S—rhyme recognition-Spanish.

Table 4
t-Tests of mean differences for observations of classroom quality

	Two-way immersion			English immersion		
	N	*M*	S.D.	*N*	*M*	S.D.
ECERS-R	20	5.11	.63	16	4.86	.65
SELA	20	3.23	.42	16	3.01	.53

ECERS-R, early childhood environmental rating scale.
SELA, supports for early literacy assessment.
t-test indicates no significant differences at .05 level.

Table 5
Post hoc sheffe tests, ANOVA for observation of support for language and culture

	TWI-Spanish			EI			TWI-English		
	N	*M*	S.D.	*N*	*M*	S.D.	*N*	*M*	S.D.
L1 use	10	3.90[a]	.32	16	1.81	.98	10	1.40	.52
Culture	10	3.50[a]	.97	16	2.25	.58	10	1.90	.99
L2 support	10	3.70	.48	16	3.25	.68	10	3.30	.68

L1 Use, the teacher speaks and uses the native language of the majority of the children in the classroom.
Culture, the teacher incorporates the cultural backgrounds and life experiences of the child into the life of the classroom.
L2 support, teacher uses effective strategies to help children understand and acquire English.
[a] Difference is significant at the .05 level compared to EI and TWI-English.

of classrooms differed from the others. Means by classrooms and results are reported in Table 5. The results indicate that teachers in the Spanish-language TWI classrooms used Spanish much more frequently than did the teachers in the other classrooms. The Spanish-language TWI classrooms also scored significantly better than the other classrooms on incorporating the cultural background into classroom life. Classrooms did not significantly differ on use of effective strategies to help children learn English. The ratings indicate that the Spanish-language TWI classrooms used Spanish much (but not all) of the time and provided relatively strong support for home culture. However, the other classrooms did provide some support for Spanish language and culture.

3.2. Children's learning

Comparison of pre-and post-tests indicates that all children made substantial learning gains (Tables 2 and 3). For both groups of children, standard scores on the PPVT increased from about 78–84, a 6 point gain. Standard score gains were roughly 5 points on the WJ-R Picture Vocabulary and 3 points on the WJ-R Applied Problems tests. Raw score gains on the unstandardized literacy measures are evident, as well. For example, children advanced from recognizing 4 or 5 letters to recognizing 10 or 11 letters in English.

Two sets of analyses were conducted to compare the effects of the TWI and EI programs, an ordinary regression with robust standard errors adjusted for the clustering of data by classroom and a hierarchical linear model (HLM). The HLM is limited by the small number of classrooms in this study, but it does allow us to account for the effects of classroom level variation in ECERS score and teacher qualifications while properly estimating standard errors at the student level. Student level variables in both analyses include age, gender, language (Spanish L1 or not), and pre-test score. Treatment was specified as a dichotomous variable (TWI = 1, EI = 0) and was a classroom level variable in the HLM. However, the intraclass correlation coefficient was not statistically significant in any of the HLM analyses. The single level regression analyses omit classroom level variables that were irrelevant in all but three of the 12 HLM analyses.

All variables entered the analysis simultaneously, and the dependent variables are raw scores. Standard scores are not available for all measures, and analyses of standard scores where available produced essentially the same results. All children in the study were included in these analyses, even those whose first language was not English ($n = 7$). Separate analyses conducted without those children yielded nearly identical results, but the analyses presented follow the "once randomized, always analyzed" approach to ensuring the integrity of the randomized design. No statistically significant interaction effects between language or gender and treatment were found. Thus, the results presented are for analyses that exclude the irrelevant interactions.

Results of the regression analyses are presented in Tables 6 and 7. Results of the HLM are presented in Tables 8 and 9. Effect size estimates from both sets of analyses are presented in Table 10. The estimated effect size was calculated as the regression coefficient divided by the control group standard deviation (Glass' delta). The estimated results of the two models are nearly identical with respect to the effects of TWI compared to EI. At conventional levels of significance ($p < .05$) no program effects were found on measures of English language and literacy development. The only hints of treatment effects on English language measures were near significant estimated advantages for the TWI program group in the ordinary regressions with robust standard errors for Applied Problems (es = .28, $p = .051$) and English Phoneme Deletion (es = .23, $p = .077$).

Table 6
Regression analysis with robust cluster standard errors estimating TWI program effects on English measures (all children)

Variable	PPVT ($n = 131$) Reg B (S.E.)	WJ-R PV ($n = 131$) Reg B (S.E.)	WJ-R AP ($n = 128$) Reg B (S.E.)	PD-E ($n = 130$) Reg B (S.E.)	Letter ID-E ($n = 131$) Reg B (S.E.)	Rhyme-E ($n = 131$) Reg B (S.E.)
TWI vs. EI	1.21 (1.74)	−.21 (.37)	1.15 (.57)	.93 (.51)	−.52 (1.30)	−.01 (.38)
Spanish L1	−2.11 (1.56)*	−.42 (.47)	.10 (.96)	−.15 (.48)	−.31 (1.16)	.07 (.31)
Age	.33 (.15)*	.04 (.03)	.08 (.05)	.18 (.05)**	.05 (.09)	.11 (.03)***
Pre-test score	.72 (.05)***	.54 (.07)***	.66 (.16)***	.13 (.11)	.91 (.07)***	.58 (.10)***
Gender	−1.13 (1.64)	−.13 (.31)	−.77 (.68)	−1.17 (.53)*	−1.70 (1.29)	−.01 (.41)
Constant	4.63 (7.20)	8.68 (1.43)***	1.19 (1.89)	1.48 (2.26)	5.12 (4.82)	−4.24 (1.32)**
R^2	.70	.71	.31	.26	.51	.39

PPVT—the Peabody picture vocabulary test.
WJ-R PV—Woodcock Johnson picture vocabulary.
WJ-R AP—Woodcock Johnson applied problems.
PD-E—Phoneme deletion-English.
Letter ID-E—alphabet recognition-English.
Rhyme-E—rhyme recognition-English.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Statistical analyses were performed on all Spanish language measures using only the sample children who were Spanish dominant. The Spanish-language children in the TWI program group gained significantly and substantially more on the TVIP (es = .61 and .56, $p < .001$) compared to their peers in the EI group. Over the school year, TVIP standard scores (i.e., age adjusted scores) improved for the TWI group, but declined for the EI group. Spanish-language children in the TWI program group also performed significantly better on the Spanish language rhyme test (es = .45, $p < .05$) in the ordinary regression. There was one hint of a possible advantage for the EI program. Spanish-language children in the EI program group scored higher than their TWI peers on Spanish letter recognition, but this effect did not reach a conventional level of significance even in the single level regression (es = −.35, $p = .065$).

In any study with multiple outcome measures, multiple statistical tests are conducted, and this increases the risk of a Type I error. One approach to taking this into account is to apply the Bonferroni correction to adjust the alpha level for statistical significance. The Bonferroni procedure is to divide the alpha level by the number of statistical tests conducted, in this case .05/12, or .004. The effect on the TVIP would be considered statistically significant. None of the other estimated treatment effects are significant if alpha is set at .004. However, the Bonferroni correction is overly conservative where the outcome measures are correlated, as is the case here.

Table 7
Regression analysis with robust cluster standard errors estimating TWI program effects on Spanish measures (Spanish-dominant children only)

Variable	TVIP ($n = 74$) B (S.E.)	WM-R PV ($n = 74$) B (S.E.)	WM-R AP ($n = 74$) B (S.E.)	PD-S ($n = 74$) B (S.E.)	Letter ID-S ($n = 74$) B (S.E.)	Rhyme-S ($n = 74$) B (S.E.)
TWI vs. EI	6.01 (1.08)***	−.01 (.56)	.21 (.62)	.84 (.60)	−3.33 (1.74)	1.08 (.47)*
Age	−.14 (.16)	−.06 (.06)	.17 (.06)**	.15 (.07)*	.27 (.10)*	.09 (.04)
Pre-test Score	.93 (.11)***	.75 (.10)***	.38 (.09)***	.24 (.09)*	.89 (.13)***	.53 (.17)**
Gender	−2.39 (1.63)	−1.65 (.69)*	−.12 (.60)	.25 (.69)	−1.71 (1.43)	−.51 (.52)
Constant	12.11 (7.43)	6.75 (3.34)	−4.89 (2.81)	1.87 (3.33)	−4.57 (4.85)	−2.90 (2.02)
R^2	.61	.50	.34	.33	.48	.28

TVIP—Test de Vocabulario en Imagenes Peabody.
WM-R PV—Woodcock-Muñoz picture vocabulary.
WM-R AP—Woodcock-Muñoz applied problems.
PD-S—Phoneme deletion-Spanish.
Letter ID-S—alphabet recognition-Spanish.
Rhyme-S—rhyme recognition-Spanish.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Table 8
HLM analysis estimating TWI program effects on English measures (all children)

Variable	PPVT ($n=131$) Coefficient (S.E.)	WJ-R PV ($n=131$) Coefficient (S.E.)	WJ-R AP ($n=128$) Coefficient (S.E.)	PD-E ($n=130$) Coefficient (S.E.)	Letter ID-E ($n=131$) Coefficient (S.E.)	Rhyme-E ($n=1\ 31$) Coefficient (S.E.)
Fixed effects						
Class-level						
Intercept	40.64 (.79)***	18.64 (.15)***	9.55 (.19)	11.43 (.25)***	10.32 (.58)***	1.53 (.16)***
ECERS	1.06 (1.09)	.72 (.23)**	.37 (.36)	−.38 (.64)	.79 (1.05)	.36 (.16)*
Master's degree	−1.23 (2.19)	−.91 (.32)**	−.56 (.36)	.58 (.57)	−1.85 (1.70)	.06 (.39)
TWI vs. EI	1.30 (1.89)	−.19 (.35)	.69 (.43)	.89 (.57)	−.18 (1.16)	−.12 (.39)
Student-level						
Spanish LI	−2.08 (1.48)	−.27 (.43)	−.48 (.53)	−.28 (.44)	−.16 (1.11)	.08 (.30)
Age	.33 (.15)*	.03 (.03)	.12 (.03)**	.18 (.05)**	.04 (.09)	.10 (.03)***
Pre-test score	.71 (.05)***	.53 (.06)***	.54 (.08)***	.14 (.10)	.89 (.08)***	.57 (.10)***
Gender	−1.08 (1.66)	−.12 (.28)	−.32 (.50)	−1.17 (.53)	−1.46 (1.18)	−.08 (.41)
Random effects	Variance	Variance	Variance	Variance	Variance	Variance
Between-class variance	4.40	.008	.003	.087	2.69	.070
Between-student variance	82.72	4.07	7.48	9.48	38.72	3.70

PPVT—the Peabody picture vocabulary test.
WJ-R PV—Woodcock Johnson picture vocabulary.
WJ-R AP—Woodcock Johnson applied problems.
PD-E—Phoneme deletion-English.
Letter ID-E—alphabet recognition-English.
Rhyme-E—rhyme recognition-English.
* $p<.05$.
** $p<.01$.
*** $p<.001$.

Table 9
HLM analysis estimating TWI program effects on English measures (Spanish-dominant children only)

Variable Fixed effects	TVIP ($n=74$) Coefficient (S.E.)	WM-R PV ($n=74$) Coefficient (S.E.)	WM-R AP ($n=74$) Coefficient (S.E.)	PD-S ($n=74$) Coefficient (S.E.)	Letter ID-S ($n=74$) Coefficient (S.E.)	Rhyme-S ($n=74$) Coefficient (S.E.)
Class-level						
Intercept	18.50 (.55)***	13.11 (.25)***	8.48 (.28)***	11.82 (.29)***	8.95 (.71)***	2.12 (.23)***
ECERS	−.18 (.95)	−.41 (.40)	.11 (.48)	.04 (.42)	2.57 (1.17)*	.59 (.36)
Master's degree	−1.41 (1.57)	1.16 (.68)	.35 (.82)	.24 (.75)	−2.97 (1.63)	−.22 (.52)
TWI vs. EI	6.58 (.94)***	−.30 (.57)	.02 (.66)	.73 (.66)	−3.19 (1.74)	.92 (.56)
Student-level						
Age	−.13 (.16)	−.07 (.06)	.17 (.06)**	.15 (.06)*	.28 (.09)**	.09 (.04)*
Pre-test score	.93 (.11)***	.75 (.09)***	.37 (.09)***	.23 (.08)**	.89 (.12)***	.53 (.16)**
Gender	−1.99 (1.59)	−1.90 (.67)**	−.02 (.63)	.18 (.80)	−1.55 (1.28)	−.55 (.49)
Random Effects	Variance	Variance	Variance	Variance	Variance	Variance
Between-class variance	.023	.005	.042	.020	4.40	.005
Between-student variance	60.50	7.45	5.68	6.62	30.78	5.62
Intraclass correlation	.001	.001	.003	.001	.060	.001

TVIP—Test de Vocabulario en Imagenes Peabody.
WM-R PV—Woodcock-Muñoz picture vocabulary.
WM-R AP—Woodcock-Muñoz applied problems.
PD-S—Phoneme deletion- Spanish.
Letter ID-S—alphabet recognition-Spanish.
Rhyme-S—rhyme recognition-Spanish.
* $p<.05$.
** $p<.01$.
*** $p<.001$.

Table 10
Estimated effect sizes for TWI compared to EI

	Effect size	
	HLM	Regression
English measures		
PPVT	.08	.07
WJ-R PV	−.05	−.05
WJ-R AP	.17	.28
PD-E	.22	.23
Letter ID-E	−.02	−.06
Rhyme-E	−.05	.00
Spanish measures		
TVIP	.61	.56
WM-R PV	−.08	.00
WM-R AP	.01	.06
PD-S	.24	.28
Letter ID-S	−.34	−.35
Rhyme-S	.38	.45

For the most part, the analyses did not find significant effects of variations in observed quality or advanced degrees on children's test scores. Statistically significant positive effects of ECERS score were found for WJ-R Picture Vocabulary, English Rhyme, and Spanish Letter Identification. Controlling for ECERS score, the estimated effect of an MA degree was negative for WJ-R Picture Vocabulary. None of these estimated classroom-level effects is significant if the Bonferoni Correction is applied.

The relative effects of TWI on Spanish-language development for English-dominant children could not be estimated in the same way because EI children with English (or another non-Spanish language) as a primary language were not tested in Spanish. However, it was possible to look at Spanish-language development for English-dominant children in the TWI classrooms over the year. Matched pairs *t*-tests were used to determine whether TWI children with English as their primary language made gains in Spanish from Fall to Spring. Statistically significant gains were found in Spanish receptive vocabulary, Spanish phoneme deletion, and Spanish letter recognition.

4. Discussion

This study used random assignment to compare the effects of TWI (Spanish and English) and EI (English) preschool education programs. Most children served were Hispanic and spoke Spanish as the primary language at the home. All children in this study made substantial progress on English language measures of vocabulary development, literacy, and mathematics over the school year. These gains compare favorably to those found for other preschool education programs with high standards and indicate that both approaches to preschool education were reasonably successful in improving children's academic skills (Barnett, Jung, Lamy, Wong, & Cook, 2007; Gormley et al., 2005). The TWI program provided better support for Spanish language receptive vocabulary development without sacrificing gains in English language development. In addition, English-dominant children in the TWI preschool program also made gains in Spanish language and literacy skills without hindering the development of their English language abilities (though this last finding is not bolstered by comparison to a control group). As discussed earlier in the paper, bilingualism is valuable *per se* and has been found to yield added benefits for children's learning and development over the long term (Hakuta & Garcia, 1989; National Association for the Education of Young Children (NAEYC), 1995).

Although the purpose of this study was to compare alternative approaches to preschool education, the data provided some basis for estimating the overall effectiveness of both approaches. Little or no standard score gain might be expected for children in the study without enrollment in a high-quality preschool education program. Quality in both programs was reasonably high as judged by structural features and observation. Standard score gains were also fairly large compared to those reported in other studies (Early et al., 2006; Zill et al., 2001). Both approaches appear to have worked well as judged by the English language measures.

We found no statistically significant indications of transfer of skills from Spanish to English. The effect sizes for English phoneme deletion might be taken as suggestive. However, several of our measures proved problematic and our outcomes are strictly short term during the preschool years. Transfer of knowledge and skills regarding the alphabet, grammar, and other aspects of language and literacy might occur later as children progress through elementary school (Bialystok, Majumder, & Martin, 2003; Durgunoglu, Nagy, & Hancin-Bhatt, 1993; English, Leafstedt, Gerber, & Villaruz, 2001; Riccio et al., 2001). This would be an important focus for follow-up with this sample and for future studies of TWI and EI preschool programs.

The strongest difference observed between programs was that Spanish-dominant children in the TWI program had much greater gains in Spanish vocabulary (receptive language) with an effect size of over half of a standard deviation. While the Spanish-dominant children in the EI program lost ground during the year relative to age norms, their peers in the TWI program gained ground relative to age norms. The decline in TVIP standard scores for Spanish-dominant children in the EI program is evidence that immersion in an English-only preschool setting is accompanied by, though it does not necessarily contribute to, Spanish language loss. The TVIP standard score gains for Spanish-dominant children in the TWI program indicate that TWI preschool programs can ameliorate this problem with no loss of effectiveness in promoting English language and literacy development.

Given this study's modest sample size, care should be taken not to conclude that failure to find additional significant differences on the majority of outcomes between treatments is strong evidence of no difference. In addition to the TVIP effect, Spanish-dominant children had significantly greater gains on the Spanish language rhyme test in the TWI program compared to the EI program in the more parsimonious single level regression. Also, there were hints of positive effects of TWI on two English language measures and one hint of a positive effect of EI. The "nearly significant" effect favoring the EI program for gains in Spanish letter identification was particularly intriguing. Letter identification is easily taught, and perhaps the EI program spent more time on letter recognition in English, Spanish, or both languages.

Given the multiple comparisons, all of the indications of effects for measures other than the TVIP might be dismissed as likely to have occurred by chance. However, we note them for two reasons. First, three of the four (including the only one significant at conventional levels) favor TWI and two of these are on English language measures. Second, these estimates suggest outcomes that should get close attention in larger future studies on the effects of TWI and other approaches on emergent literacy skills in English and Spanish (and other languages).

Despite the strong gains for the TWI group in Spanish receptive vocabulary, comparable gains were not found for Spanish expressive vocabulary. Both groups experienced declines in standard scores and raw scores on the WM-R Picture Vocabulary test. We found no indication that the TWI program contributed to improved performance in Spanish expressive language compared to the EI program. The striking difference in outcomes between expressive and receptive vocabulary indicates that research on approaches to educating ELL students should take care to measure both. It would be useful for future research to investigate why such differences occur and how positive outcomes might be extended to expressive language.

A limitation of this study is that the emergent literacy skills tests used were new, data on their performance is limited, and they did not all perform as well as we would have liked. It is possible that impacts on phonological awareness and letter naming would have been more evident had we employed better measures. Future research would benefit from further development of comparable measures in English and Spanish (and other languages) for the language and literacy development of young children. This study sought to measure growth comparably in both languages, and we have encountered several limitations that others also have noted (NAEYC, 2005; Pearson, Fernandez, & Oller, 1993; Restrapo & Siverman, 2001; Tabors et al., 2003). In our experience, many children from low-income language minority families fail to get any items correct at age three on commonly used measures. This problem was found for one of the new measures we employed. It has since been revised and refined. Better measures for young children will increase what can be learned from studies in this vital area of first and second language and literacy development.

In addition, future research could benefit from detailed qualitative descriptions of specific within-program literacy supports to obtain more information about the details of program implementation including, for example, the degree to which each program emphasizes letter knowledge and other specific features of literacy support. Such information could help interpret both overall results and specific variations in findings that might not universally favor one approach.

We did not conduct a detailed cost analysis or benefit-cost analysis of the two programs. If offered for the same number of hours, the TWI program would not require additional resources. It accomplishes its goals by hiring native

Spanish speakers as lead teachers for half of the preschool classrooms rather than by adding additional staff to existing classrooms. However, in our study, the TWI program operated for somewhat more hours and days than the EI program, even though both were encompassed by wrap-around child care that extended time to the same number of hours and days. We cannot separate effects of the additional time from other aspects of the program design and have no way of knowing if the TWI program would have been just as effective without extra hours to deliver the program. This is an important question for further study.

The study has several other notable limitations. It is modest in size and duration and limited to one particular location and population. Results might differ for children with other languages or even for Spanish-dominant children with different backgrounds. The number of English-dominant children was small, and the number of children with a third primary language was too small for separate analyses. In addition, the treatments were not pure. By design, the TWI program was 100% immersion in each language on alternate weeks and the EI program was English language immersion with some support for first language development. The differences between programs were large, but neither model was "pure" immersion. The divergence from pure immersion models is to be expected (Stipek et al., 2001) and may have been adaptive or may have detracted from the effectiveness of one or both of the models. Future research can address these and other limitations.

No single study should be expected to provide an adequate basis for policy and practice. This study provided some added insights into the effectiveness of TWI and EI when the preschool education is of reasonably high quality. The study's methodology provides confidence that the findings accurately represent program effects. The TWI preschool approach supported stronger Spanish language development at no detriment, and possible benefit, to English language development, literacy, and math achievement of both native English and native Spanish speakers. The EI preschool approach was similarly effective in promoting English language development, literacy, and math skills. Both approaches employed highly qualified teachers regardless of language used in the classroom. All of our findings are limited to immediate program impacts. Nevertheless, the findings should reassure policymakers that high quality preschool education using either a TWI approach or EI can substantially benefit ELL children. There may be other benefits to the children from increased bilingual proficiency that we did not identify because we did not assess broader cognitive outcomes or social and emotional outcomes and the study continued only through the end of a single year of preschool education (Portes & Hao, 2002).

Our findings indicate that it would be useful to continue to develop and evaluate TWI approaches to preschool education. Future randomized trials could investigate how effectiveness may vary with specific program characteristics, populations served, community and school contexts, and other conditions that might affect success. TWI has the potential to address two serious problems—inadequate progress in native language for the nation's rapidly growing ELL population, and inadequate second language acquisition of native English speakers. The need for improvement in these areas is great and their importance grows daily. Success in these domains might lead to additional long-term success in attaining other educational goals. Further studies of preschool TWI and EI approaches can help to produce and refine practical solutions to these educational problems.

References

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

Barnett, W. S., Hustedt, J. T., Robin, K. B., & Schulman, K. L. (2005). *The state of preschool: 2005 state preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research, Rutgers University.

Barnett, W. S., Jung, K., Lamy, C., Wong, V., & Cook, T. (2007). *Effects of five state prekindergarten programs on early learning*. Paper presented at the SRCD annual meeting, Boston, MA.

Bialystok, E., & Herman, H. (1999). Does bilingualism matter for early literacy? *Bilingualism: Language and Cognition*, *2*, 35–44.

Bialystok, E. (2006). Second language acquisition and bilingualism at an early age and the impact on early cognitive development. In R. E. Trembley, R. G. Barr, & R. DeV. Peters (Eds.), Encyclopedia on Early Childhood Development [online]. Retrieved January 9, 2007 from http://www.excellence-earlychildhood.ca/documents/BialystokANGxp.pdf.

Bialystok, E., Majumder, S., & Martin, M. (2003). Developing phonological awareness: Is there a bilingual advantage? *Applied Psycholinguistics*, *24*, 27–44.

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, *7*, 325–339.

Blanco, D. (2004). *Alphabet recognition test*. New Brunswick, NJ: National Institute for Early Education Research.

Campbell, R., & Sais, E. (1995). Accelerated metalinguistic (phonological) awareness in bilingual children. *British Journal of Developmental Psychology*, *13*, 61–68.

Campos, S. (1991). The Carpinteria preschool program: A long-term effects study. In E. Garcia & B. McLaughlin (Eds.), *Meeting the challenge of linguistic and cultural diversity in early education* (pp. 34–48). New York: Teachers' College Press.

Collins, R., & Ribeiro, R. (2004). Toward an early care and education agenda for Hispanic children. *Early Childhood Research & Practice*, 6(2). Retrieved April 14, 2005 from http://ecrp.uiuc.edu/v6n2/collins.html.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *California State Department of Education. Schooling and language minority students: A theoretical framework* (pp. 3–49). Los Angeles, CA: California State University, Evaluation, Dissemination and Assessment Center.

DeKeyser, R. M., & Larson-Hall, J. (2005). What does the critical period really mean? In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 89–108). Oxford: Oxford University Press.

Diaz, R., Padilla, K., & Weathersby, E. (1991). The effects of bilingualism on preschoolers' private speech. *Early Childhood Research Quarterly*, *6*, 377–393.

Dickinson, D., & Tabors, P. (Eds.). (2001). *Beginning literacy with language: Young children learning at home and school*. Baltimore, MD: Paul H. Brooks.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test Third edition (PPVT-III)*. Circle Pines, MN: American Guidance Service.

Dunn, L., Lugo, D., Padilla, E., & Dunn, L. (1986). *Test de Vocabulario en Imagenes Peabody*. Circle Pines, MN: American Guidance Service.

Durgunoglu, A. Y., Nagy, W. E., & Hancin-Bhatt, B. J. (1993). Cross-language transfer of phonological awareness. *Journal of Educational Psychology*, *85*, 453–465.

Early, D., Bryant, D., Pianta, R., Clifford, R., Burchinal, M., Ritchie, S., et al. (2006). Are teachers' education, major, and credentials related to classroom quality and children's academic gains in pre-kindergarten? *Early Childhood Research Quarterly*, *21*, 174–195.

English, J. P., Leafstedt, J., Gerber, M. M., & Villaruz, J. (2001). *Individual differences in phonological skills for Spanish speaking kindergartners learning English: Relationship between English and Spanish phonological measures*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Frede, E. C., & Barnett, W. S. (1992). Developmentally appropriate public school preschool: A study of implementation of the High/Scope Curriculum and its effects on disadvantaged children's skills at first grade. *Early Childhood Research Quarterly*, *7*, 483–499.

Garcia, G. (2000). Bilingual children's reading. In M. Kamil, P. Mosenthal, P. Pearson, & R. Barr (Eds.), *Handbook of reading research* (pp. 813–834). Mahwah, NJ: Erlbaum.

Gormley, W. T., Gayer, T., Phillips, D., & Dawson, B. (2005). The effects of universal pre-k on cognitive development. *Developmental Psychology*, *41*, 872–884.

Greene, J. P. (1997). A meta-analysis of the Rossell and Baker review of bilingual education research. *Bilingual Research Journal*, *2*(2/3). Retrieved June 7, 2006, from. http://brj.asu.edu/articlesv2/green.html

Hakuta, K. (1986). *Mirror of language: The debate of bilingualism*. New York, NY: Basic Books.

Hakuta, K., & Garcia, E. (1989). Bilingualism and education. *American Psychologist*, *44*, 374–379.

Harms, T., Clifford, R., & Cryer, D. (1998). *Early Childhood Environmental Rating Scale-Revised*. New York: Teacher's College Press.

Head Start National Reporting System (2006). Results of Fall 2005 NRS assessments. Presentation of the Head Start National Reporting System, March 15, 2006. Retrieved June 19, 2006 from http://www.hsnrs.net/Fall2005ResultsWebCast_031506.ppt.

Hohmann, M., & Weikart, D. P. (1995). *Educating young children: Active learning practices for preschool and child care programs.* Ypsilanti, MI: High/Scope Press.

Howard, E. R., Sugarman, J., & Christian, D. (2003). *Trends in two-way immersion education: A review of the research*. Baltimore, MD: CRESPAR/Johns Hopkins University. [Report 63]

Huempfner, L. (2004). Can one size fit all? The imperfect assumptions of parallel achievement tests for bilingual students. *Bilingual Research Journal*, *28*, 379–399.

Johnson, D., & Walker, T. (1991). A follow-up evaluation of the Houston Parent-Child Development Center: School performance. *Journal of Early Intervention*, *15*, 226–236.

Johnson, J. S., & Newport, E. I. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.

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

Lamy, C. (2004). *Measuring early literacy: Assessing children and classrooms to improve practice*. Presentation at the International Reading Association's annual conference, Reno, NV.

Lamy, C., Frede, E., Seplocha, H., Strasser, J., Jambunathan, S., Juncker, J., et al. (2004). *Giant steps for the littlest children: Progress in the sixth year of the Abbott Preschool Program Year three initial update 2004–2005*. Trenton: New Jersey Department of Education.

Lindholm-Leary, K. (2005). *Review of research and best practices on effective features of dual language education programs*. San Jose, CA: San Jose State University.

Lopez, M. (2005). *Following the development of phonological awareness skills in low-income bilingual children from preschool through kindergarten*. Poster presented at the Biennial Meeting of the Society for Research in Child Development, Atlanta, GA.

Lopez, L. M., & Greenfield, D. B. (2004). The cross-language transfer of phonological skills of Hispanic Head Start children. *Bilingual Research Journal*, *28*, 1–18.

National Association for the Education of Young Children (NAEYC). (1995). *Responding to linguistic and cultural diversity: Recommendations for effective early childhood education. A position statement of the National Association for the Education of Young Children*. Washington, DC: Author.

National Association for the Education of Young Children (NAEYC). (2005). *Screening and assessment of young English-language learners: Supplement of the NAEYC position statement on early childhood curriculum, assessment, and program evaluation*. Washington, DC: Author.

National Center for Early Development and Learning (NCEDL) (2005). Pre-kindergarten in eleven states: NCEDL's multi-state study of prekindergarten & study of state-wide early education programs (SWEEP). Preliminary descriptive report. Retrieved August 5, 2005 from http://www.fpg.unc.edu/~ncedl/pages/products.cfm#sweep_ms.

National Institute for Early Education Research (NIEER). (2005). *Support for English language learners classroom assessment*. New Brunswick, NJ: Author.

Pearson, B. Z., Fernandez, S. C., & Oller, D. K. (1993). Lexical development in bilingual infants and toddlers: Comparison to monolingual norms. *Language Learning*, *43*, 93–120.

Peisner-Feinberg, E., & Burchinal, M. (1997). Relations between preschool children's child care experiences and concurrent development: The cost, quality, and outcomes study. *Merrill-Palmer Quarterly*, *43*, 451–477.

Portes, A., & Hao, L. (2002). The price of uniformity: Language, family, and personality adjustment in the immigrant second generation. *Ethnic and Racial Studies*, *25*, 889–912.

Puma, M., Bell, S., Cook, R., Heid, C., Lopez, M., Zill, N., et al. (2005). *Head Start impact study: First year findings*. Washington, DC: US Department of Health and Human Services, Administration for Children and Families.

Reese, L., Garnier, H., Gallimore, R., & Goldenberg, C. (2000). A longitudinal analysis of the ecocultural antecedents of emergent Spanish literacy and subsequent English reading achievement of Spanish-speaking students. *American Educational Research Journal*, *37*, 633–662.

Restrapo, M., & Siverman, S. (2001). Validity of the Spanish preschool language scale-3 for use with bilingual children. *American Journal of Speech Language Pathology*, *10*, 382–393.

Reynolds, A. J. (1991). *Bilingualism, multiculturalism, and second language learning*. Hillsdale, NJ: Erlbaum.

Riccio, C. A., Amado, A., Jimenez, S., Hasbrouck, J. E., Imhoff, B., & Denton, C. (2001). Cross-linguistic transfer of phonological processing: Development of a measure of phonological processing in Spanish. *Bilingual Research Journal*, 25(4) [Electronic version]. Retrieved June 7, 2006 from http://brj.asu.edu/content/vol25_no4/pdf/ar9.pdf.

Rodriguez, J. L., Diaz, R. M., Duran, D., & Espinosa, L. (1995). The impact of bilingual preschool education on the language development of Spanish-speaking children. *Early Childhood Research Quarterly*, *10*, 475–490.

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

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.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 77–121). Timonium, MD: York Press.

Schweinhart, L. J., Montie, J., Xiang, Z., Barnett, W. S., Belfield, C. R., & Nores, M. (2005). *Lifetime effects: The High/Scope Perry Preschool study through age 40. (Monographs of the High/Scope Educational Research Foundation, 14)*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Short, D. (1993). *Integrating language and culture in middle school American history classes*. Santa Cruz, CA and Washington, DC: National Center for Research on Cultural Diversity and Second Language Learning. [Educational Practice Rep. No. 8]

Slavin, R. E., & Cheung, A. (2005). A synthesis of research on language of reading instruction for English language learners. *Review of Educational Research*, *75*, 247–281.

Smith, M. S. (1975). Evaluation findings in Head Start Planned Variation models. In A. M. Rivlin & P. M. Timpane (Eds.), *Planned variation in education* (pp. 101–111). Washington, DC: The Brookings Institution.

Smith, S., Davidson, S., Weisenfeld, G., & Katsaros, S. (2001). *Supports for early literacy assessment (SELA)*. New York: New York University.

Smith, T., Kleiner A., Parsad, B., & Farris, E. (2003). *Prekindergarten in U.S. public schools: 2000–2001 (NCES 2003-019)*. Washington, DC: US Department of Education, National Center for Educational Statistics. Retrieved June 7, 2006 from http://nces.ed.gov/pubs2003/2003019.pdf.

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

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

Sylva, K., & Siraj-Blatchford, I. (2001). *The relationship between children's developmental progress in the pre-school period and two rating scales*. Paper presented at the International ECERS Network Workshop, Santiago, Chile.

Sylva, K., Siraj-Blatchford, I., Taggart, B., Sammons, P., Melhuish, E., Elliot, K., & Totsika, V. (2006). Capturing quality in early childhood through environmental rating scales. *Early Childhood Research Quarterly*, *21*, 76–92.

Tabors, P. O., Paez, M. M., & Lopez, L. M. (2003). Dual language abilities of bilingual four-year olds: Initial findings from the Early childhood study of language and literacy development of Spanish-speaking children. *NABE Journal of Research and Practice*, *1*, 70–91.

Tabors, P. O., & Snow, C. E. (2001). Young bilingual children and early literacy development. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy* (pp. 159–178). New York, NY: Guilford Press.

U.S. Census Bureau (2000). Census 2000 summary file3 (SF 3): Tables P145A-P1451. Retrieved April 15, 2005 from http://factfinder.census.gov/servlet/DatasetTableListServlet?_ds_name=DEC_2000_SF3_U&_type=table&lang=en&_program=DEC&ts=7564142654.

Whitehurst, G. J., & Lonigan, C. J. (2001). Emergent literacy: Development from prereaders to readers. In S. B. Neuman & D. K. Dickinson (Eds.), *Handbook of early literacy research* (pp. 11–29). New York, NY: Guilford Press.

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

Winsler, A., Diaz, R., Espinosa, L., & Rodriguez, J. (1999). When learning a second language does not mean losing the first: Bilingual language development in low-income. Spanish-speaking children attending bilingual preschool. *Child Development*, *70*, 349–362.

Woodcock, R., & Johnson, M. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Allen, TX: DLM Teaching Resources.

Woodcock, R., & Munoz-Sandoval, A. F. (1996). *Bateria Woodcock-Munoz: Pruebas de habilidad cognitive-Revisada*. Chicago, IL: Riverside Publishing.

Wong Filmore, L. (1991). When learning a second language means losing the first. *Early Childhood Research Quarterly*, *6*, 323–346.

Zill, N., Resnick, G., Kim, K., McKey, R., Clark, C., Pai-Samant, S., et al. (2001). *Head Start FACES: Longitudinal findings on program performance. Third progress report*. Washington, DC: U.S. Department of Health and Human Services.