Three Theories of the Effects of Language Education Programs: An Empirical Evaluation of Bilingual and English-Only Policies

JEFF MACSWAN
University of Maryland

MARILYN S. THOMPSON
Arizona State University

KELLIE ROLSTAD
University of Maryland

KARA McALISTER
Arizona State University

GERDA LOBO
Arizona State University

ABSTRACT

We empirically evaluated three theoretical models—the threshold hypothesis, transfer theory, and time-on-task theory—for educating English language learners (ELLs), with a focus on the role of language factors in explaining achievement differences among ELLs. Participants were 196 sixth graders with Spanish language backgrounds who started learning English in kindergarten and then were continuously enrolled in a U.S. school. Structural equation modeling was used to estimate the extent to which Spanish and English language and literacy—skills that are emphasized differentially in competing theories for educating ELLs—predict academic achievement assessed in English. Results indicated that Spanish literacy, over and above English language proficiency, was substantially predictive of academic achievement, consistent with the transfer theory. This model was a more focused version of the threshold hypothesis, in that the weaker predictor of Spanish oral language proficiency was excluded. Time-on-task theory was not supported.

INTRODUCTION

Young second language (L2) learners in U.S. schools, or bilingual learners, represent a substantial proportion of the school-age population. Known in policy contexts as English language learners (ELLs), these students represented 9.3% of all U.S. public school children in 2013–2014. Between 6% and 9.9% of students were ELLs in Arizona, Arkansas, Delaware, Florida, Hawaii, Illinois, Kansas, Maryland, Massachusetts, Minnesota, New York, North Carolina, Oklahoma, Oregon,
Rhode Island, Virginia, and Washington; whereas Alaska, California, Colorado, Nevada, New Mexico, and Texas had ELL enrollments of 10% or more. California’s ELL population was the largest, at 22.7%. Most of these students (76.5%) spoke Spanish as a first language. The National Assessment of Educational Progress has documented a substantial achievement gap between ELLs and non-ELLs since data collection began in the 1970s, and this gap persists to the present day. This achievement gap underscores the special challenges ELLs face in school. (For more details, see Kena et al., 2016.)

In a landmark decision, *Lau v. Nichols* (1974), the Supreme Court unanimously found that a “disparate impact” occurred when a San Francisco school district failed to provide for the special challenges facing school-age L2 learners, as “students who do not know English are effectively foreclosed from any meaningful education.” The Court found that the policy violated the Civil Rights Act of 1964, and that the school district must provide students with “appropriate relief.” School districts have responded differently to *Lau* as well as other important court decisions and legislation. Some have implemented intensive programs to teach English and maximize children’s exposure to English, such as structured English immersion (SEI), while others use children’s first language (L1) as a support to help them keep up academically in school subjects while learning English.

A number of studies have been carried out to evaluate the relative effectiveness of different educational programs for bilingual learners. These studies have established a strong consensus in the field favoring L1 support, a program approach known as bilingual education, but relatively little empirical research has been done to evaluate specific theories underlying the positive effects of L1 support. Moreover, English-only approaches to educating ELLs persist in most states. In this article, we seek to move beyond the program effectiveness literature to identify the extent to which first and second language and literacy—emphasized differently in various approaches to educating ELLs—explain achievement differences among ELLs. In particular, we employ a structural equation modeling (SEM) approach to empirically evaluate three theoretical models proposed to explain academic achievement differences among ELLs—the threshold hypothesis, transfer theory, and the time-on-task theory—with a focus on the role language plays in such theories.

**Research on Language Minority Education**

*Program Effectiveness Research: Description*

Slavin and Cheung (2005) conducted a “best evidence” synthesis of research on bilingual education programs focusing on methods of teaching reading to ELL students, using methods typical of narrative reviews. They compared academic outcomes between ELL students taught using a bilingual approach (with the L1 used before or simultaneously with the L2 in reading instruction) and those taught using an SEI approach (with reading instruction in English only). Slavin and Cheung’s review, based on a corpus of 16 studies, concluded that most
methodologically rigorous studies favored bilingual approaches over English-only approaches; although some did not find a difference, none favored English-only approaches over bilingual approaches. Thus, Slavin and Cheung concluded that the available evidence favored bilingual approaches, especially paired bilingual strategies that teach reading in the native language and English at the same time.

While most other narrative reviews (August & Hakuta, 1997; August & Shanahan, 2006; Fillmore & Valadez, 1986; Genesee, Lindholm-Leary, Saunders, & Christian, 2006; Meyer & Fienberg, 1992; National Academies of Sciences, Engineering, and Medicine, 2017) have reached similar conclusions, two reviews co-authored by Keith Baker (Baker & de Kanter, 1981; Rossell & Baker, 1996) were more critical of bilingual education and remain highly influential among advocates of English-intensive programs. While the Baker reviews are often cited as favoring English-only approaches, the authors’ own conclusions state only that policy should not promote an exclusive reliance on bilingual education, but should support the use of an SEI approach as well (Baker & de Kanter, 1981, chap. 4, p. 6; Rossell & Baker, 1996, p. 19). The Baker reviews have been criticized for including studies that did not meet the authors’ stated selection criteria, for assigning multiple “votes” to studies published in different forms, and for misclassifying program models (Greene, 1998; Krashen, 1996; McField & McField, 2014; Rolstad, Mahoney, & Glass, 2005; Reljić, Ferring, & Martin, 2015; Slavin & Cheung, 2005).

In addition to narrative reviews, several authors have also conducted meta-analyses of the effectiveness of language education programs for ELLs. For instance, Willig (1985), Greene (1998), and Rolstad et al. (2005) found bilingual education to be superior to English-only approaches, and Rolstad et al. additionally discovered that developmental bilingual education programs, which continue to support children’s L1 long after they develop proficiency in English, were even more effective at promoting children’s academic achievement than transitional, early-exit bilingual education programs, which emphasize transitioning to all-English instruction as soon as possible. Rolstad et al. reported that studies controlling for ELL status indicated a positive effect for bilingual education of 0.23 standard deviation, with outcome measures in the native language showing a positive effect of 0.86 standard deviation. More recently, McField and McField (2014) conducted a meta-analysis of all previous meta-analyses of bilingual education in the United States, and Reljić et al. (2015) conducted a meta-analysis of bilingual education programs in Europe, modeled after Rolstad et al., both drawing similar conclusions regarding the effectiveness of bilingual approaches. Overall, this extensive body of meta-analytic research on the effectiveness of bilingual education programs has consistently found the use of the home language at school to be more effective than English alone at promoting ELL children’s academic achievement.

Program Effectiveness Research: Theory

While considerable research has been done on the effectiveness of bilingual education programs, little empirical research has been done to evaluate specific theories
### Type of Bilingualism

<table>
<thead>
<tr>
<th>Type of Bilingualism</th>
<th>Cognitive Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Additive bilingualism.</strong> High levels in both languages</td>
<td>Positive cognitive effects</td>
</tr>
<tr>
<td><strong>Dominant bilingualism.</strong> Native-like level in one of the languages</td>
<td>Neither positive nor negative cognitive effects</td>
</tr>
<tr>
<td><strong>Semilingualism.</strong> Low level in both languages (may be balanced or dominant)</td>
<td>Negative cognitive effects</td>
</tr>
</tbody>
</table>

**Figure 1.** The Threshold Hypothesis (adapted from Cummins, 1979, p. 230).

about the underlying causes of the positive effects associated with the use of children’s home language in instructional settings. In the present study, we empirically evaluate three such theories—the threshold hypothesis, transfer theory, and the time-on-task theory.

The threshold hypothesis (Cummins, 1976, 1979, 2000) is a leading theory of academic achievement differences among bilingual students. It posits that language minority children undergo native language loss, or a shift in language dominance from the L1 to the L2, and that “the level of linguistic competence attained by a bilingual child may mediate the effects of his bilingual learning experiences on cognitive growth.” Cummins (1979, 1981) included both literacy and language in his conception of “low ability in both languages.” While it will doubtless be true that children with low English literacy will also have low academic achievement, essentially by definition, the more provocative claim that Cummins made is that linguistic minorities in the United States are characterized by “less than native-like command of the vocabulary and syntactic structures of both L1 and L2” (Cummins, 1979, p. 238), and that variation in L1 oral language is indicative of “low ability” in the L1 (Cummins, 1994, p. 3814). Cummins (1979, p. 230) presented the threshold hypothesis graphically (Figure 1).

MacSwan (2000) noted a concern with the threshold hypothesis in that it posits that some children may have limited (oral) proficiency in their L1 while proficiency in the L2 is still developing, a condition Cummins and others (Hansegård, 1968; Ringbom, 1962; Skutnabb-Kangas, 1981; Toukomaa & Skutnabb-Kangas, 1977) have called “semilingualism.” The existence of such a condition was never
established empirically, and a review of the evidence has shown that it does not occur in typically developing bilingual children (Commins & Miramontes, 1989; Edelsky, 2006; Edelsky et al., 1983; Hakuta & D’Andrea, 1992; MacSwan, 2000; MacSwan & Rolstad, 2003, 2006, 2010; Martin-Jones & Romaine, 1986; Paulston, 1983; Petrovic & Olmstead, 2001; Valadez, MacSwan, & Martínez, 2000; Wiley & Rolstad, 2014; Wiley, 2005).

Different facets of language proficiency are included in Cummins’s conception of bilingualism as embedded in the threshold hypothesis. Our analysis will therefore examine various aspects of language (use) referenced by Cummins, together and separately, and attempt to determine what variability exists among learners and whether these variations in abilities relate to success at school, as the threshold hypothesis posits. By unpacking language proficiency into discrete components of oral language and literacy, and evaluating the effects of each on academic achievement, we hope to gain a more nuanced view of the hypothesized causal relationships illustrated in Figure 1.

As a complement to the threshold hypothesis, Cummins (1981, p. 29) also theorized about the advantage of bilingual education in terms of transfer theory in the linguistic interdependence hypothesis:

To the extent that instruction in \( L_x \) \([\text{Language } x]\) is effective in promoting proficiency in \( L_x \), transfer of this proficiency to \( L_y \) \([\text{Language } y]\) will occur provided there is adequate exposure to \( L_y \) (either in school or in the environment) and adequate motivation to learn \( L_y \).

To illuminate, Cummins (1981) contrasted the perspective of a common underlying proficiency with a separate underlying proficiency model; the common underlying proficiency model, which Cummins endorsed, relates separate “surface features” of both languages to a common underlying proficiency which is shared between them, facilitating transfer. As Cummins more recently elaborated,

In concrete terms, what this hypothesis implies is that in, for example, a dual language Spanish-English bilingual program in the United States, Spanish instruction that develops Spanish reading and writing skills is not just developing Spanish skills, it is also developing a deeper conceptual and linguistic proficiency that is strongly related to the development of literacy in the majority language (English). In other words, although the surface aspects (e.g., pronunciation, fluency, etc.) of different languages can be distinguished, there is an underlying cognitive/academic proficiency that is common across languages. This common underlying proficiency makes possible the transfer of cognitive/academic or literacy-related proficiency from one language to another. (Cummins, 2017, p. 106)

Krashen (1996) similarly emphasized the role of transfer across languages, but without specifically drawing upon the linguistic interdependence hypothesis. For Krashen, bilingual instruction reduces achievement differences between bilingual and monolingual English-speaking children because it allows ELLs to keep up academically while they struggle to learn and understand English; this idea was
the basic rationale for bilingual education programs noted by the United States Commission on Civil Rights (1975) and embedded in *Lau v. Nichols* (1974). Krashen argued that language minority children in all-English instruction who achieve well are able to do so because they are likely to have educational resources at home that provide a kind of “de facto bilingual education”; that is, although these children cannot understand academic instruction at school because they do not know English, they receive L1 support for school subject learning at home. In addition, Krashen argued, exposure to academic content in L1 allows children to achieve higher levels of proficiency in L2 because it provides a network of contextual clues for children to make inferences regarding the meaning of new words and linguistic structures presented in English. There is substantial empirical evidence that knowledge of school subjects, and literacy in particular, transfers across children’s two languages (for summaries, see August & Shanahan, 2006; Cummins, 2017; Genesee et al., 2006; National Academies of Sciences, Engineering, and Medicine, 2017).

As MacSwan and Rolstad (2005) have noted, a concern with the linguistic interdependence hypothesis is the embedded assumption that language and language-related academic content matter are not distinct. The surface features of the L1 and L2 for Cummins (1981, p. 25) are those that “have become relatively automatized or less cognitively demanding, whereas the underlying proficiency is that involved in cognitively demanding communicative tasks,” that is, in-school talk developed around academic tasks. If language proficiency is understood to include aspects of school knowledge, as in Cummins’s model, then it follows that growth in academic subjects will result from instruction in the L1. Cummins thus captures the transfer effect in his model of bilingualism, but the model is plagued by a problem: Pedagogical content knowledge is characterized as part of our “underlying cognitive/academic proficiency that is common across languages” rather than as discretely represented, context-dependent conceptual knowledge. As such, the linguistic interdependence hypothesis implies that language developed around nonschool contexts, typically used by lower socioeconomic communities (e.g., African American English and working-class English), is less developed than the language variety used at school. This aspect of the linguistic interdependence hypothesis makes the theory vulnerable to some of the same criticisms that have been levied against the threshold hypothesis (MacSwan, 2000).

MacSwan and Rolstad (2005) developed an approach to transfer theory that specifically differentiates between language and conceptual understanding of school subject matter, drawing on neurocognitive research around psychological modularity. In this view, language is cognitively special and discrete; all typically developing children acquire the language of their speech community, effortlessly and without instruction, shaped by the specific contexts in which it is acquired as an “accidental product of varied experience” (Chomsky, 1995, p. 6). Language clearly links to the conceptual knowledge developed in school, but knowledge of school subjects and other topics is distinct from linguistic knowledge itself and not part of our linguistic development. (For summaries of relevant psycholinguistic research, see Curtiss, 2013; Gallistel & King, 2009.) Viewed in these terms, transfer
is simply a metaphor for the accessibility of conceptual knowledge through the various languages people may know. Because transfer occurs across languages in this way, learning school subjects in an L1 develops conceptual knowledge available to children as they develop proficiency in their L2.

This conception of the cognitive underpinnings of transfer is consistent with the view that multilingual speakers have an underlying integrated language system, with both shared and discrete linguistic resources (MacSwan, 2017). For instructional purposes, languages might be used separately or together, using codeswitching, translation, or other syncretic processes, depending on a teacher’s objectives, as many have suggested (Cook, 2001; Cummins, 2008, 2017; Faltis, 1989; García, 2009; Jacobson, 1978, 1990; Milk, 1986). Content knowledge is not specifically associated with one language or the other, but is gained and accessed through either.

The study reported here specifically evaluates this version of transfer theory rather than the similar but somewhat differently conceptualized linguistic interdependence hypothesis.

Finally, Rossell and Baker (1996) proposed that “time-on-task” in L2 is the central factor underlying achievement among ELLs. According to their time-on-task theory, the more time children spend hearing, speaking, or studying English the higher their level of L2 proficiency will be, which in turn will translate into higher scores on academic achievement measures. As Rossell and Baker (1996, p. 22) put it, “The ‘time-on-task’ principle [is] … the notion that the amount of time spent learning a subject is the greatest predictor of achievement in that subject.” Porter (1990), also an advocate of English-only instruction, defined the time-on-task principle in this way: “The more time spent learning a language, the better you do in it, all other factors being equal” (p. 119). Like the linguistic interdependence hypothesis, the time-on-task theory does not distinguish between learning a language and learning school content in a language. The difference is that the linguistic interdependence hypothesis posits the existence of a common underlying proficiency as a repository of some aspects of school content knowledge, whereas the time-on-task theory does not.

Each of the theories described recommends a treatment aimed at achieving a specific intermediate effect, which in turn is hypothesized to predict academic achievement outcomes in English. The threshold hypothesis and transfer theory both recommend bilingual education as an instructional strategy. For the threshold hypothesis, the hypothesized intermediate cognitive effect is increased overall bilingualism (understood to be language proficiency, inclusive of both literacy and oral language in both L1 and L2), with the predicted result that participants will achieve higher academic achievement in English. The transfer theory, on the other hand, hypothesizes that bilingual education will have an intermediate effect of increasing academic achievement in the L1 as participants concurrently learn English, with the predicted result that children will transfer this knowledge to the L2 environment. The time-on-task theory, by contrast, recommends that children spend as much time as possible in English, with the intermediate effect of increased English language proficiency and predicted result of higher academic achievement in English. These relationships are summarized in Table 1.
We may thus think of the transfer theory as a more focused version of the threshold hypothesis. Transfer theory looks to improved L1 literacy, a component of academic achievement that is developed concurrently with English language proficiency, as an intermediate effect. The threshold hypothesis, on the other hand, posits that both L1 oral language and L1 literacy are related to increased academic achievement in the L2.

The underlying causes of student success are clearly much more complex than any of these models propose. The degree to which the recommended treatment is likely to achieve the intermediate effect will no doubt be moderated by numerous factors besides language of instruction, such as teacher quality, school resources, instructional methods, and more (August & Hakuta, 1997; Genesee et al., 2006; National Academies of Sciences, Engineering, and Medicine, 2017). However, as previously noted, there are clear overall effects stemming from language of instruction; here we focus narrowly on this effect, evaluating the link between a hypothesized intermediate effect and the predicted outcome to inform the merits and utility of each of these three theories as explanatory models. For Spanish academic achievement measures, we focus on Spanish literacy.

METHODS

Research Questions

We empirically examine the merits of each of the three theories by addressing the following research questions:

1. **Threshold hypothesis**: Do both Spanish language proficiency (Spanish language and literacy) and English language proficiency (English language and literacy) predict academic achievement measured in English?
2. **Transfer theory**: Does Spanish literacy, in particular, contribute over and above English language proficiency to the prediction of academic achievement measured in English?
3. **Time-on-task theory**: Is English language proficiency sufficient as a predictor of academic achievement measured in English, reflecting English-only approaches of the time-on-task theory, or is prediction of achievement improved by inclusion of Spanish language and Spanish literacy?

**Participants**

Study participants came from seven school districts in a major metropolitan area. To be included in the sample, participants had to be Spanish language-background sixth graders at the time of the study who had been continuously enrolled in a U.S. school since kindergarten, and who had not started learning English until kindergarten. Participants were educated in a variety of minority language educational programs, including English as a second language (ESL), bilingual education, dual language immersion, and SEI. For data collection, each year over a 3-year period, we obtained a list of all ELLs and reclassified ELLs entering the sixth grade in all participating school districts. We sent letters of invitation home; the letters included a brief survey to determine whether the participants met the selection criteria. Of those who responded, 196 children were identified as eligible to participate in the study.

The assessments noted in the next section were administered in clusters for each language, with at least 2 weeks between the matching Spanish and English versions of tests to avoid priming effects.

**Instrumentation**

**Language and Literacy Measures.** Using standard methods in the study of child language, students were asked to interact with native speakers of English and Spanish on separate occasions spanning 4–6 weeks and tell a story about a boy and a frog from Mercer Mayer’s picture book with no text (Mayer, 1969). These speech samples were coded for lexical, morphological, and syntactic structures and errors.

Speech samples of each child telling the whole story depicted in the picture book were individually videotaped. These speech samples were then transcribed word for word and coded using MacWhinney’s (1995) standard CHAT format, as modified by Curtiss, MacSwan, Schaeffer, Kural, and Sano (2004) and adapted to Spanish by Valadez et al. (2000). An example of a coded Spanish utterance is presented next, with translation provided in brackets:

\*MAR: El niño se está durmiendo, y la rana se escapó.

[The boy is going to sleep, and the frog escaped.]

%mor: DART|el niño REF|se IAUX|está-3Ss dormir-DUR conj|y DART|la rana REF|se IF|escapar-pret-3Ss

%lex: N|niño N|dormir N|rana N|escapar

All transcripts were coded by native speakers, and each coded transcript was proofed by at least one other native speaker trained in the coding system. When judgments regarding error coding differed, these were resolved using procedures.
detailed in Curtiss et al. (2004). In this coding system, errors of selection (e.g., when *la* is used where *el* is required for the morphological category DART) were prefixed with = (equal sign); errors of omission (a category such as DART or IAUX is missing altogether) were suffixed with =0. Erroneous lexical selection was similarly noted on the %lex: (lexical) tier, and errors in word order, if present, were noted with appropriate annotations on a %syn: (syntactic) tier. This system permitted the calculation of morphological and syntactic error rates in English and Spanish for the children in the study. Because morphological structure is regarded as central to grammatical representation, both historically in the study of child language acquisition (Brown, 1973) and in contemporary linguistic theory (Chomsky, 1995; Uriagereka, 2012), we focused our analysis on morphological error rate as an index of participants’ knowledge of grammatical structure in English and Spanish.

In addition, we administered the *Woodcock Language Proficiency Battery* (WLPB) in English and Spanish to assess oral language proficiency and literacy. Literacy measures included the Basic Reading Skills and the Reading Comprehension Clusters of the WLPB in English and Spanish, which focus on passage comprehension, vocabulary, phonics skills, and structural analysis (Woodcock, 1991). These two composite scores together were preferred over the Broad Reading composite, as they assess a wide range of reading skills. (In addition, because Broad Reading shares subscales that are used in the Basic and Comprehension composites, including the Broad Reading composite in models with the other composites would have been statistically inappropriate due to overlapping scales across the composite.) Productive measures of vocabulary were collected using the *Clinical Evaluation of Language Fundamentals* (CELF, in English and Spanish; Semel, Wiig, & Secord, 1995) in English and Spanish, and receptive vocabulary knowledge was tested using the *Peabody Picture Vocabulary Test* (PPVT; English) and *Test de Vocabulario en Imágenes Peabody* (PVIP; Spanish).

Of the oral language measures, we regard the language samples to be very solid measures of children’s knowledge of grammatical structure in English and Spanish. Although a standard stimulus was used to elicit the samples, the context of the task provided a broad opportunity for participants to demonstrate their knowledge of the grammatical structure of English and Spanish. The English Oral Language Cluster of the WLPB, on the other hand, is focused on English used in academic contexts and may measure language proficiency concurrently with aspects of academic achievement (MacSwan & Pray, 2005). Similarly, we note that the *Peabody* appears similarly to focus on school-related vocabulary and may have substantial construct overlap with academic achievement, unlike the CELF, an open-ended and more context-independent vocabulary measure. We therefore regard the natural language samples and CELF to be good indicators of language proficiency and believe the *Woodcock Oral Language measure* and the *Peabody* vocabulary test to be more narrowly representative of language used in the context of schooling.

Language measures were collected in both Spanish and English for all children in the study. Key constructs and their measures are summarized in Table 2. In the analyses, we initially consider oral language and literacy as separate constructs,
with oral language including measures of grammatical structure as well as vocabulary.

**Academic Achievement Measures.** Academic achievement was measured using the Reading, Mathematics, and Language scores of the *Stanford Achievement Test, 9th Edition* (SAT-9), as administered at the end of the sixth grade and obtained from students’ academic files. As is evident from the data summaries and analyses presented in the next sections, participants in the study appear to have had sufficiently well-developed English language proficiency to minimize measurement error introduced by limited proficiency in the linguistic medium of the test (for discussion, see Thompson, DiCerbo, Mahoney, & MacSwan, 2002).

**Analysis**

A structural equation modeling (SEM) framework was used to examine the hypothesized relationships between English and Spanish language proficiency and academic achievement as assessed in English. All analyses were conducted using full-information maximum likelihood estimation within Mplus 6.11, which accommodates data missing at random. Preliminary analyses were conducted to examine measures descriptively, including computations of means and standard deviations for all English and Spanish language measures and for the academic achievement outcomes, as well as correlations among the measures.

The modeling process consisted first of an evaluation of measurement models relating the hypothesized constructs to the language and achievement measures, followed by the estimation of latent regression models in which the constructs underlying English and Spanish language proficiency predict academic achievement. The global fit of each model was evaluated according to the model chi-square statistic and several goodness-of-fit indices, including the comparative fit index (CFI), root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). The Akaike information criterion (AIC) was also included as a parsimony-corrected index to support comparisons of fit across nonnested models; smaller values indicate better fit. The proportions of variance in the academic achievement factor explained by the latent regression models were also examined using $R^2$ indices.

<table>
<thead>
<tr>
<th>Construct</th>
<th>Measures in Both Versions: Spanish (L1) and English (L2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral Language Proficiency</td>
<td></td>
</tr>
<tr>
<td>Grammar</td>
<td>Language sample (frog story)</td>
</tr>
<tr>
<td>Vocabulary</td>
<td><em>Clinical Evaluation of Language Fundamentals</em> (CELF) <em>Peabody</em> (Spanish PVIP; English PPVT)</td>
</tr>
<tr>
<td></td>
<td><em>Woodcock Oral Language Cluster</em></td>
</tr>
<tr>
<td>Literacy</td>
<td><em>Woodcock Reading Comprehension</em></td>
</tr>
<tr>
<td></td>
<td><em>Woodcock Basic Reading Skills</em></td>
</tr>
</tbody>
</table>

Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0267190517000137
Three alternative measurement models were first fit to the data. These measurement models differed in the number and nature of factors used to model covariances among the six proficiency measures within each language: (a) with separate factors for literacy, school-based language, and natural language; (b) with separate factors for literacy and oral language development (i.e., combining school-based and natural language into a single factor); and (c) with a single language proficiency factor. It was expected that the language factors would be strongly correlated within both Spanish and English, so although a nuanced view of the structure of language was desired, we had concerns about encountering multicollinearity problems in the latent regression model in which these correlated factors predicted academic achievement. Finally, latent regression models were then specified in which academic achievement was predicted from the English and Spanish language factors supported in the measurement models.

RESULTS

Preliminary Analyses

Based on an analysis of multivariate outliers, data for five students were omitted from further analyses due to high Mahalanobis distances, influence statistics, and Cook’s distances. The resulting data set for analysis consisted of 191 cases. Percentages of cases with missing data for the measures ranged from less than 2% for the Spanish Peabody to approximately 20% for SAT-9 achievement measures. The major reasons for missing observations were absence during administration for select English and Spanish language measures and, for reasons unknown and unavailable to us, missing SAT-9 scores in the school data files. Estimates of correlations, means, and standard deviations for the measures are reported in Table 3, as computed in the full-information maximum likelihood procedure used in all SEM analyses.

As expected, the language measures most strongly correlated with the three academic achievement measures (i.e., reading, language, mathematics) were the English reading comprehension, basic reading, and oral/vocabulary language measures; these correlations all exceeded .5 except for the one between mathematics and English basic reading. Additionally, several measures of Spanish literacy and language were substantially correlated with the English achievement outcomes. Correlations between Spanish reading comprehension and the three achievement measures ranged from .39 (mathematics) to .47 (reading); Spanish basic reading, oral/vocabulary, and the Peabody were also correlated with achievement outcomes in the .3 to .4 range. In contrast, the language sample measures (i.e., frog story and CELF) were minimally correlated with the achievement outcomes, with the exception of the English CELF’s correlation with reading (.39) and mathematics (.44). Scores on the frog story tended to be particularly high and to display little variance, indicating students in the sample were proficient speakers of both languages (rates of correct morphological responses: English $M = 97.75, SD = 1.35$;
## TABLE 3. Correlations and Descriptive Statistics Among Measures

<table>
<thead>
<tr>
<th>Measures</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Sp. Comp.</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Sp. Basic</td>
<td>0.657</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Sp. Peabody</td>
<td>0.480</td>
<td>0.286</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Sp. WOral</td>
<td>0.757</td>
<td>0.464</td>
<td>0.543</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Sp. Frog</td>
<td>0.226</td>
<td>0.308</td>
<td>0.169</td>
<td>0.208</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Sp. CELF</td>
<td>0.393</td>
<td>0.171</td>
<td>0.256</td>
<td>0.381</td>
<td>0.081</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Eng. Comp</td>
<td>0.464</td>
<td>0.314</td>
<td>0.377</td>
<td>0.420</td>
<td>0.048</td>
<td>0.198</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Eng. Basic</td>
<td>0.526</td>
<td>0.616</td>
<td>0.278</td>
<td>0.373</td>
<td>0.267</td>
<td>0.111</td>
<td>0.650</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Eng. Peabody</td>
<td>0.234</td>
<td>0.077</td>
<td>0.377</td>
<td>0.299</td>
<td>0.068</td>
<td>0.105</td>
<td>0.582</td>
<td>0.379</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Eng. WOral</td>
<td>0.305</td>
<td>0.100</td>
<td>0.414</td>
<td>0.365</td>
<td>0.012</td>
<td>0.128</td>
<td>0.807</td>
<td>0.496</td>
<td>0.666</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11. Eng. Frog</td>
<td>0.031</td>
<td>0.039</td>
<td>0.066</td>
<td>0.005</td>
<td>0.025</td>
<td>0.035</td>
<td>0.371</td>
<td>0.258</td>
<td>0.318</td>
<td>0.352</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12. Eng. CELF</td>
<td>0.221</td>
<td>0.018</td>
<td>0.127</td>
<td>0.210</td>
<td>0.111</td>
<td>0.354</td>
<td>0.502</td>
<td>0.227</td>
<td>0.370</td>
<td>0.542</td>
<td>0.208</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13. Language</td>
<td>0.419</td>
<td>0.388</td>
<td>0.277</td>
<td>0.306</td>
<td>0.128</td>
<td>0.024</td>
<td>0.546</td>
<td>0.549</td>
<td>0.363</td>
<td>0.501</td>
<td>0.181</td>
<td>0.283</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14. Math</td>
<td>0.390</td>
<td>0.303</td>
<td>0.306</td>
<td>0.299</td>
<td>0.126</td>
<td>0.106</td>
<td>0.533</td>
<td>0.388</td>
<td>0.419</td>
<td>0.528</td>
<td>0.147</td>
<td>0.438</td>
<td>0.690</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>15. Reading</td>
<td>0.471</td>
<td>0.361</td>
<td>0.368</td>
<td>0.414</td>
<td>0.213</td>
<td>0.102</td>
<td>0.615</td>
<td>0.507</td>
<td>0.505</td>
<td>0.579</td>
<td>0.224</td>
<td>0.370</td>
<td>0.820</td>
<td>0.687</td>
<td>1.000</td>
</tr>
</tbody>
</table>

*Note.* Sp. = Spanish; Comp. = Comprehensive Reading; WOral = Woodcock Oral; CELF = Clinical Evaluation of Language Fundamentals; Eng. = English; SD = standard deviation.
Table 4. Global Fit Indices for Measurement Models (CFAs) and Latent Variable Regression Models

<table>
<thead>
<tr>
<th>Model Description</th>
<th>$\chi^2$</th>
<th>df</th>
<th>CFI</th>
<th>SRMR</th>
<th>RMSEA [90% CI]</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Seven-factor CFA, with academic achievement in Eng. and literacy, school-based language, and natural language factors within both Eng. and Sp.</td>
<td>130.51</td>
<td>67</td>
<td>.95</td>
<td>.12</td>
<td>.070 [.052, .088]</td>
<td>17,444.68</td>
</tr>
<tr>
<td>2. Five-factor CFA, with school-based and natural language combined into an oral language factor within both Eng. and Sp.</td>
<td>143.43</td>
<td>78</td>
<td>.95</td>
<td>.121</td>
<td>.066 [.049, .083]</td>
<td>17,435.60</td>
</tr>
<tr>
<td>3. Four-factor CFA; same as M2 except with a single Eng. proficiency factor (combining literacy, school-based language, and natural language)</td>
<td>186.80</td>
<td>82</td>
<td>.92</td>
<td>.16</td>
<td>.082 [.066, .097]</td>
<td>17,470.97</td>
</tr>
<tr>
<td>4. Latent regression predicting achievement from Eng. proficiency, Sp. literacy, and Sp. oral language; statistically equivalent to M3</td>
<td>186.80</td>
<td>82</td>
<td>.92</td>
<td>.16</td>
<td>.082 [.066, .097]</td>
<td>17,470.97</td>
</tr>
<tr>
<td>5. Latent regression nested within M4; path from Sp. oral language to achievement set to 0</td>
<td>188.45</td>
<td>83</td>
<td>.92</td>
<td>.16</td>
<td>.082 [.066, .097]</td>
<td>17,470.617</td>
</tr>
<tr>
<td>6. Latent regression nested within M4 and M5, paths from Sp. oral language and Sp. literacy to achievement set to 0</td>
<td>188.45</td>
<td>83</td>
<td>.92</td>
<td>.16</td>
<td>.082 [.066, .097]</td>
<td>17,470.617</td>
</tr>
</tbody>
</table>

Note. df = model degrees of freedom; CFI = comparative fit index; SRMR = standardized root mean square residual; RMSEA = root mean square error of approximation; CI = confidence interval; AIC = Akaike information criterion; Eng. = English; Sp. = Spanish; CFA = confirmatory factor analysis; M = model.

Spanish $M = 98.46, SD = 1.39$). Accordingly, this restriction of range limits the correlation of the frog story with other measures.

Measurement Models

Alternative measurement models (i.e., confirmatory factor analysis [CFA] models) were fit to the data to examine the hypothesized relationships between the language and achievement measures; adequate measurement models are a necessary precursor to specifying structural relations in a latent regression analysis. Global fit statistics for measurement and latent variable regression models are presented in Table 4. First, a seven-factor CFA was specified that included an academic
achievement factor and separate factors for literacy, school-based language, and natural language in English and Spanish underlying the six language measures within each language. Construct overlap between corresponding measures in different languages is a plausible source of covariance not accounted for adequately by the factor structure, so we examined whether it was necessary to allow all respective measures to covary between languages. Based on modification indices and statistical significance of the estimated parameters across different models, only between-language error covariances for basic reading and the CELF were deemed necessary and were included in this and all subsequent models. This seven-factor model (Model 1) fit the data well, $\chi^2 (67) = 130.51, p < .001$, RMSEA = .07, 90% CI [.05, .09], CFI = .95, SRMR = .12. All factors were significantly ($p < .001$) and positively correlated with academic achievement except Spanish natural language ($r = .28, p = .12$): English literacy (.72), English school language (.66), English natural language (.78), Spanish literacy (.53), and Spanish school language (.49). Additionally, correlations among language factors within Spanish and within English were statistically significant and exceeded .85; estimated correlations of English school-based language with English natural language and English literacy were slightly outside of bounds (>1.0) within this model.

To investigate the need for maintaining separate factors for school and natural language measures within each language, a second CFA was specified that included separate factors for literacy and oral language development within each language (i.e., school-based and natural language loaded on a single oral language factor). This five-factor model (Model 2) fit the data similarly well, $\chi^2 (78) = 143.43, p < .001$, RMSEA = .07, 90% CI [.05, .08], CFI = .95, SRMR = .12. All factors were significantly correlated with academic achievement ($p < .001$): English literacy (.71), English oral language (.66), Spanish literacy (.53), and Spanish oral language (.47).

Given the very high correlation between English literacy and oral proficiency (.91), as well as concerns about encountering multicollinearity problems in the latent regression models, one additional CFA was specified in which a single factor of English proficiency underlay the six English language measures. Both literacy and oral language factors were maintained for the Spanish measures given the focal goal of estimating separate effects for these factors on academic achievement. The fit of this four-factor model (Model 3) was marginally adequate, $\chi^2 (82) = 186.80, p < .001$, RMSEA = .08, 90% CI [.06, .10], CFI = .92, SRMR = .16.

**Latent Regression Models**

Latent regressions were specified to estimate the unique effects of the various language factors on academic achievement, reflecting the three theories being evaluated. As expected, when latent regressions were estimated using multiple-factor structures of both English and Spanish language, multicollinearity resulting from high correlations among language and literacy factors within each language caused instability in the regression coefficients (even changing signs) and inflated standard errors. To minimize these difficulties, the highly correlated English oral
language and literacy factors were combined into a single English language proficiency factor (as in Model 3) in the latent regression analyses.

Model 4 was estimated to address the threshold hypothesis, in which Spanish language proficiency (Spanish language and literacy) and English language proficiency (English language and literacy) predict academic achievement measured in English. Because all paths were freely estimated, the fit of Model 4 was equivalent to its respective measurement model (Model 3). Figure 2 shows partial regression coefficients for Model 4 (first set of estimates); Spanish literacy ($\beta_1 = .50$) and English language proficiency ($\beta_3 = .64$) contributed positively and significantly to the prediction of academic achievement, whereas Spanish oral language ($\beta_2 = -.30$) did not contribute to academic achievement after controlling for Spanish literacy and English language proficiency. In the CFA (Model 3), however, Spanish oral language was correlated positively with academic achievement ($r = .47$), so this negative, nonsignificant estimated coefficient may be attributable to multicollinearity due to the substantial correlation between

---

**Figure 2.** Standardized Path Coefficients for Latent Regression Models Predicting Academic Achievement. The two sets of partial regression coefficients reflect coefficients for Models 4 and 5, respectively; the path from L1 oral language was constrained to 0 in Model 5.
Spanish literacy and oral language \((r = .88)\). The three predictors accounted for 58\% of the variance in the academic achievement factor, which dropped to 56\% when Spanish oral language was excluded (Model 5).

Models 5 and 6, respectively, sequentially constrained to 0 the paths from L1 oral language \((\beta_2)\) and L1 literacy \((\beta_1)\) to academic achievement; both models had similar, marginally adequate fit to the data. To comparatively evaluate both the transfer theory and the time-on-task theory, the contribution of Spanish literacy to academic achievement over and above English language proficiency was examined. A chi-square difference test comparing Model 5 to Model 6 was statistically significant, \(\Delta \chi^2 (1) = 10.27, p < .01\), indicating that eliminating the path from Spanish literacy to academic achievement significantly worsened the fit of the model and decreased the \(R^2\) from 56\% (Model 5) to 52\% (Model 6), supporting the positive contribution of Spanish literacy to the prediction of academic achievement beyond that contributed by English language proficiency. In Model 5 (second set of coefficients on Figure 2), the standardized partial regression coefficients for Spanish literacy \((\beta_1 = .24)\) and English language proficiency \((\beta_3 = .60)\) were both statistically significant.

**Discussion and Conclusions**

Our model results indicated that Spanish literacy together with English language proficiency are substantially predictive of academic achievement in English, consistent with the transfer theory. Our examination of transfer theory was based on evaluating the contribution of Spanish literacy to achievement, controlling for English language proficiency. We noted that this model was a more focused version of the threshold hypothesis, in that the weaker predictor of Spanish oral language proficiency was excluded. Model effect sizes were nearly as large as for the more complex threshold hypothesis.

Upon a closer look at both the correlations of measures with achievement scores as well as regression coefficients, we found that the Spanish language sample and the Spanish CELF vocabulary measure were not substantially correlated with academic achievement. In comparison, the Spanish literacy and English oral proficiency measures and, to a lesser extent, the more school-oriented Spanish oral proficiency measures (Peabody and Woodcock Oral Language Cluster) were more strongly positively related to academic achievement.

The time-on-task theory emphasizes L2 (English) language proficiency at the cost of instructional time spent learning in the L1 (Spanish). Our findings did not support this approach. In particular, Spanish literacy significantly predicted academic achievement measured in English over and above English language proficiency \((R^2\) increase of 4\% over the 52\% of variance accounted for by English proficiency). We conducted an additional analysis that included Spanish literacy as a sole predictor of academic achievement and found that it accounted for 29\% of the academic achievement in English. So, students would be well served academically by a language program that supports their growth in literacy in L1 while learning L2.
A long-standing controversy in language minority education, with important consequences for curriculum and instruction for ELLs, is the nature of bilingualism and language growth in immigrant children and how their bilingualism may or may not be related to school success. An outcome of the conception of proficiency embedded within the threshold hypothesis and related notions has been the widespread belief among teachers and educational researchers that ELLs may enter into a state of “semilingualism” (Cummins, 1979) or “limited bilingualism” (Cummins, 1981), in which children are said to know neither their L1 nor their L2 with native-like proficiency. As has been pointed out previously (Edelsky, 2006; Edelsky et al., 1983; MacSwan, 2000; MacSwan & Rolstad, 2003, 2006, 2010; Rolstad, 2015; Wiley, 2005; Wiley & Rolstad, 2014), talk of such linguistic hierarchies stands to negatively impact children and their communities in unintended ways, and these assertions lack the empirical and theoretical support one should expect.

The transfer theory, however, avoids these pitfalls. This theory, a traditional notion in bilingual education policy and embedded in Krashen’s (1996) work as well as Cummins’s (1981, 2008, 2017) linguistic interdependence hypothesis, suggests that academic subject knowledge transfers and indeed facilitates academic development in the L2 environment. Here, we do not point to differences in the native language ability of ELLs (and hence to that of their families and communities) as a factor predicting school success, but rather focus on elements indigenous to school culture itself, namely, the learning of literacy and other school-based subjects. We prefer to refer to this model as the transfer theory rather than the linguistic interdependence hypothesis because we believe that the latter erroneously equates language development with academic content knowledge development, while transfer theory (for us) does not.

We have further shown that the time-on-task theory, which seeks to provide a rationale for English-only and English-focused programs, is not supported by our empirical findings, consistent with program evaluation reviews (McField & McField, 2014; Reljić et al., 2015; Rolstad et al., 2005; Slavin & Cheung, 2005). The core conceptual problem with the time-on-task theory is its failure to distinguish between academic content learning, independent of a linguistic medium, and the linguistic medium itself. Once this distinction is properly understood, the time-on-task principle—that “the amount of time spent learning a subject is the greatest predictor of achievement in that subject” (Rossell & Baker, 1996, p. 22)—is entirely consistent with the transfer theory, and is supportive of bilingual education: The amount of time spent learning academic subjects may be the greatest predictor of achievement in those subjects, but time spent learning will be meaningless if it is not spent in a language children can understand.

LIMITATIONS

We note two limitations of our study. First, we were not able to track student program placements in English-only and bilingual education programs consistently due to inadequate documentation and unanticipated student program changes. We
addressed this limitation by focusing on the hypothesized intermediate effects of different program types noted in Table 1. It would be interesting in future studies to examine relationships between particular program features and academic outcomes. Second, our student participants were drawn from schools across seven school districts. Accordingly, the data would be regarded as clustered to some degree due to multiple students being sampled from particular schools that were nested within these districts. Although this violates the assumption of independence in sampling, we did not have sufficient numbers of school or district clusters to account statistically for this nonindependence in the data structure. It is possible that the standard errors of our parameter estimates are slightly smaller than might be expected if data were sampled randomly, which can increase the chance of finding significant effects.

IMPLICATIONS

Theories are important because they provide a conceptual framework from which a wide range of inferences may be drawn. Our present study sought to gain insight into competing theories of academic achievement differences for bilingual students precisely because different theories hold very different sets of implications for educational policy and practice. For instance, the threshold hypothesis may imply for some that instruction should focus specifically on language-related aspects of L1 development, attending to vocabulary and grammar in the L1 rather than using the L1 as a medium for teaching and learning in content areas. Further, the threshold hypothesis may suggest for some that L1 development needs to precede L2 teaching, with the goal of achieving a developmental threshold in the L1 before the L2 can effectively become a focus of instruction. The threshold hypothesis presupposes that the language children bring with them to school, and hence the language of their community, is hierarchically related to the language of schooling and is in need of further development to facilitate school success. The time-on-task theory, on the other hand, which stems from an entirely different understanding of the basic descriptive facts, implies that teachers of ELLs should devote as much time as possible to English-mediated instruction, making little or no use of the L1 as a medium of instruction.

Transfer theory, in contrast to both the threshold hypothesis and the time-on-task theory, conceptualizes the L1 as a vehicle by which content knowledge is conveyed to learners and is not itself in need of improvement or development. Literacy is a domain of language use specific to the school setting and does not represent a further stage of linguistic development, as is commonly believed. This perspective views school as a specific context for language use and does not treat school language as a special language variety corresponding to special cognitive abilities that are inaccessible to language users in out-of-school contexts (MacSwan & Rolstad, 2003, 2005, 2010; Rolstad, 2015). Development of children’s home language is not a prerequisite to English language learning, so the teaching of ESL may proceed concurrently and independently. Thus, bilingual education contributes to
children’s success at school because it provides them with access to content area knowledge so they can keep up academically during the time it takes them to learn English.

The concept of second language instructional competence (SLIC) serves as a complement to transfer theory (MacSwan & Rolstad, 2003; Rolstad, 2005, 2015; Rolstad & MacSwan, 2008). Offered as an alternative to Cummins’s Basic Interpersonal Communication Skills/Cognitive-Academic Language Proficiency (BICS/CALP) dichotomy, SLIC refers to children’s development of L2 proficiency, not to proficiency in their home or community language; once children have learned English well enough to understand school subject matter instruction in English, they have developed SLIC. Second language acquisition is also facilitated as a result of L1 instruction because, as Krashen (1996) argued, it strengthens children’s knowledge of the context for L2 acquisition, contributing to the comprehensibility of L2 input (on the role of background knowledge in L2 acquisition, see Bygate, 2001; Gass & Varonis, 1984). Thus, transfer theory views children’s home language as a critically important resource that allows access to comprehensible content area instruction mediated through the home language with the concurrent teaching of ESL. As children’s knowledge of school content grows, it promotes the acquisition of English by creating a more effective context for L2 teaching and learning, with the goal of gradually achieving SLIC when children can meaningfully and effectively engage with English-mediated instruction.

From a policy perspective, transfer theory suggests that children’s home language should be actively used as a resource for the purposes of teaching and learning. It contrasts with the threshold hypothesis, which posits a developmental relationship between the home language and the L2, and with the time-on-task theory, which emphasizes English-mediated instruction and views use of children’s home language in classroom settings as an essential waste of instructional time and effort. Transfer theory views children’s emerging bilingualism holistically (Grosjean, 1985, 2010), permitting and encouraging access to their full linguistic repertoire (García, 2009; MacSwan, 2017) as an ongoing resource to aid their success in school.

Acknowledgments

We gratefully acknowledge research support from the U.S. Department of Education Institute of Education Science, award # R305T000164. We are grateful for the support and assistance of students, teachers, administrators, and undergraduate and graduate student research assistants, especially Jessie Ortiz.

NOTES

1. English-only programs continue to predominate in school systems across the United States despite research clearly favoring first language support for a variety of political and practical reasons. See Crawford (2004, 2008) for discussion.
2. For a detailed description of program models, see Baker and Wright (2017).
REFERENCES


doi:10.1002/9781444310498.ch13


