HHS Public Access

Author manuscript

Biling (Camb Engl). Author manuscript; available in PMC 2023 January 13.

Published in final edited form as:

Biling (Camb Engl). 2022 November ; 25(5): 899–912. doi:10.1017/s1366728922000104.

Characterization of English and Spanish language proficiency among middle school English learners with reading difficulties

Kelly T. Macdonald[1], **David J. Francis**[1], **Arturo E. Hernandez**[1], **Anny P. Castilla-Earls**[2], **Paul T. Cirino**[1]

[1]Department of Psychology, Texas Institute for Measurement, Evaluation, and Statistics, University of Houston, Houston, Texas

[2]Department of Communication Disorders and Sciences, University of Houston, Houston, Texas

Abstract

Among bilinguals, language-related variables such as first and second language PROFICIENCY and BALANCE may be related to important cognitive and academic outcomes, but approaches to characterizing these variables are inconsistent, particularly among at-risk samples of children. The current study employed comprehensive language assessment of English and Spanish language skills and contrasted various approaches to the characterization of language among at-risk ELs in middle school ($N = 161$). Specifically, we contrasted variable-centered and person-centered approaches, and convergence between objective and self-report measures. Findings support a two-factor structure of English and Spanish language skills in this population, three profiles of students (balanced, moderately unbalanced-higher Spanish, and very unbalanced-higher English), convergence between variable-centered and person-centered approaches, and mixed support for subjective indices of usage. Results provide a foundation from which to examine the roles of L1 and L2 proficiency as well as balance in important cognitive and academic outcomes in this at-risk and understudied population.

Keywords

English learners; balanced bilingualism; language; proficiency

Introduction

The proportion of the population that speaks a language other than English at home increased by 148% in the United States between 1980 and 2009, totaling 57.1 million people, or 20% of the population over the age of 5 (Ortman & Shin, 2011). Increasing linguistic diversity has led to a rise in bilingual research across the domains of psychology, neuroscience, and education. While much bilingual research has focused on bilinguals as a group, such generalization likely obscures variability due to individual differences. Two important sources of this variability are: 1) language PROFICIENCY LEVELS (e.g., as measured

Address for correspondence: Kelly Macdonald, M.A., University of Houston, Department of Psychology, Texas Institute for Measurement, Evaluation, and Statistics (TIMES) Health 1 Building, 4849 Calhoun, Room 477 Houston, TX 77204, kmacdonald@uh.edu.

by objective standardized language assessments) in both the first (L1) and second (L2) languages; and 2) the extent to which bilinguals are BALANCED in their proficiency levels. Both of these factors have been associated with cognitive and academic outcomes (Kim, Lambert & Burts, 2018; Sandhofer & Uchikoshi, 2013).

It is particularly relevant to consider these sources of bilingual variability among populations at risk for adverse outcomes, as such knowledge may inform identification and intervention approaches for those needing it most. In addition, there are likely benefits conferred by bilingualism, and it is especially important to recognize these potentially protective factors among at-risk groups. However, the bilingual literature is inconsistent with regard to methods used for characterizing language proficiency levels and balance, and few studies have done so in samples of at-risk children. Thus, the purpose of this study is to compare various approaches to the characterization of language proficiency levels and balance in middle school English Learners (ELs), an at-risk population of bilingual children.

ELs are at higher risk for academic difficulties since they must work to become proficient in English in addition to learning subject material (Hammer, Jia & Uchikoshi, 2011; Hoff, 2013; National Center for Educational Statistics, 2003). ELs from low socioeconomic backgrounds who attend under-resourced urban schools are at even higher risk for adverse outcomes including school failure and attrition (Bradley & Corwyn, 2002). In this context, risk is further exacerbated when ELs have identified difficulties in reading (Francis, Rivera, Lesaux, Kieffer & Rivera, 2006). Importantly, most work evaluating language and achievement among bilingual students is focused on younger children, whereas less is known about these relationships at the middle school level. This is particularly important given the critical importance of middle school achievement in predicting outcomes such as graduation rates and employment status (Balfanz, Herzog & Mac Iver, 2007). In the United States, the majority of ELs from low socioeconomic backgrounds speak Spanish as a first language and English as a second language, and this population is growing rapidly (Halle, Hair, Wandner, McNamara & Chien, 2012; Passel, Cohn & Lopez, 2011). While it is also important to clarify the roles of English and Spanish language processes in academic outcomes in this population, doing so presupposes a strong methodological framework for characterizing language. Thus, the primary focus of the current study is on evaluating/comparing various approaches to the characterization of language.

Dimensionality and measurement of language

When considering the measurement of L1 and L2 skills, it is important to understand how language measures relate to, versus differentiate from, one another. Although a unified view of language has been proposed (Goodman, 1997; MacWhinney, 2008), language is more often conceptualized as multidimensional (Bloom & Lahey, 1978; Pinker, 1998). Whereas a unified view posits that language may be represented as a single construct (regardless of how it is measured), a multidimensional view makes distinctions according to language demands (i.e., expressive vs. receptive language; American Psychiatric Association, 2013; CELF-5, Wiig, Secord & Semel, 2013) or specific language skills (i.e., semantic knowledge vs. syntactics; Bloom & Lahey, 1978; Pinker, 1998).

Empirical evidence from factor analytic studies provides support for a unified view of language in younger monolingual children, with increasing multidimensionality of semantics versus syntax as children get older (Foorman, Koon, Petscher, Mitchell & Truckenmiller, 2015; Lonigan & Milburn, 2017; Tomblin & Zhang, 2006). For example, a seminal study from Tomblin and Zhang (2006) found that a model which differentiated semantics from syntax provided only slight, non-significant improvement over the unitary models – with the exception of 8th grade, where the two-factor model had a significantly better fit.

Available bilingual studies evaluating the structure of language in children are limited but support a multidimensional view with a distinction between English and Spanish skills (Gottardo & Mueller, 2009; Language and Reading Research Consortium (LARRC), Yeomans-Maldonado, Bengochea & Mesa, 2018). For example, Gottardo and Mueller (2009) evaluated language and word reading skills in both Spanish and English in a sample of 1st and 2nd grade ELs. The best fitting model included two oral language factors, one English and one Spanish, each composed of both semantic and syntactic measures. There are few studies that examine these relationships among at-risk bilingual children, however, and none with older children. Thus, a better understanding of the structure of language in the specific at-risk context of middle school ELs who are also struggling readers is needed to inform measurement and characterization of language proficiency and balance. Importantly, in this population (middle school ELs with reading difficulties) we expect language levels to fall below normative expectations.

Language proficiency and balance in bilinguals

Approaches used to assess proficiency and balance are quite varied. Not only is there no gold standard in the literature for how to classify bilinguals in terms of both proficiency and balance; but, of the work that has been done, few studies address at-risk samples of children, and few utilize/compare multiple approaches. In reviewing this literature, there is a lack of clarity with regard to five issues in particular: (1) the use of self-report vs. objective metrics; (2) the specific TYPES of measures used; (3) the approach used to define BALANCE; (4) the context of the sample; and (5) whether characterization of bilinguals encompasses BOTH balance and proficiency.

The first issue is that studies index language using objective measures (i.e., Archila-Suerte, Woods, Chiarello & Hernandez, 2018; Lonigan, Goodrich & Farver, 2018; Rosselli, Ardila, Lalwani & Vélez-Uribe, 2016), self-report measures (i.e., Anderson, Mak, Chahi & Bialystok, 2018; Kim et al., 2018; Li, Sepanski & Zhao, 2006; Yow & Li, 2015), or a combination of the two (i.e., Bedore, Peña, Summers, Boerger, Resendiz, Greene, Bohman & Gillam, 2012; Gollan, Weissberger, Runnqvist, Montoya & Cera, 2012; Sheng, Lu & Gollan, 2014). Across studies that consider both approaches, correlations are moderate and generally range from $r = .40$ to $r = .60$. Self-report measures are often utilized as a proxy for language proficiency; key arguments for the use of these measures are that they are less time-consuming, require fewer resources to administer, provide more contextual information (e.g., language usage preferences across contexts), and can be easily adapted for use across a wide range of languages. In addition, some objective bilingual assessments utilize self-report

items prior to administration of test items to inform which language appears dominant in the child, then administration proceeds in the dominant language (Martin, 2013). Self-report (or parent report) regarding language usage and/or proficiency may also be used at the beginning of a comprehensive assessment to inform the language of the evaluation. Thus, it is important to better understand the extent to which such ratings converge with objective measures, particularly in this at-risk population.

A second, related issue is wide variability with regard to the specific TYPES of measures that are used; for example, within studies that utilize self-report measures, some assess language usage in different contexts (i.e., Kim et al., 2018) whereas others assess perceived language proficiency level (Gollan et al., 2012; Marian, Blumenfeld & Kaushanskaya, 2007; Sheng et al., 2014); still others consider both (i.e., Anderson et al., 2018; Li et al., 2006, 2014; Yow & Li, 2015). However, the extent to which these types of ratings relate to objective language assessments is unclear, particularly in at-risk populations. There is also variability regarding specific types of objective tests used across studies. Specifically, some studies use a single test across languages (often a measure of picture naming; Gollan et al., 2012; Sheng et al., 2014), whereas others use multiple tests and compute composite scores (i.e., Archila-Suerte et al., 2018; Rosselli et al., 2016). The present study considers a wide range of objective language assessments in both English and Spanish which are also considered alongside a self-report measure of language usage.

A third issue that emerges in reviewing these studies is the approach used to define BALANCE, with some studies utilizing continuous approaches such as factor scores (i.e., Anderson et al., 2018), difference scores (i.e., Yow & Li, 2015) or other metrics/formulas (i.e., Gollan et al., 2012; Vaughn & Hernandez, 2018). In contrast, other studies create sub-groupings of students through latent profile analysis (i.e., Kim et al., 2018; Lonigan et al., 2018), median or mean splits (i.e., Archila-Suerte et al., 2018; Rosselli et al., 2016) or other cut-off scores (i.e., Vega & Fernandez, 2011). In the present study, we utilize both variable-centered (factor analysis) and person-centered (latent profile analysis) approaches to characterize language, and then evaluate the extent to which results converge with a continuous metric of balance as well as a self-report measure of usage.

The fourth issue is that sample context varies along dimensions of age, risk status, and languages spoken, among others. Most studies are with adults, with fewer focused on children (Archila-Suerte et al., 2018; Bedore et al., 2012; Kim et al., 2018; Lonigan et al., 2018; Sheng et al., 2014), and even fewer with children identified as at-risk (Kim et al., 2018; Lonigan et al., 2018). Moreover, not all of the aforementioned studies utilize samples of Spanish–English speaking bilinguals (Anderson et al., 2018; Sheng et al., 2014; Yow & Li, 2015). The present study focuses on Spanish–English speaking middle school ELs with reading difficulties.

A fifth important issue is whether studies distinguish between language proficiency and balance. Many studies focus on degree of balanced bilingualism; that is, the extent to which the individual knows each language equally well or whether one language is stronger than the other. Although the term “balanced bilingualism” is often used in such studies, others may prefer the term “language dominance.” While it is possible for an individual to have

roughly equivalent proficiency in their L1 and L2 (a within-person distinction), they may, particularly in at-risk contexts, have low proficiency in one or both languages relative to expected norms for language skills (a between-person distinction). These within-person and between-person distinctions are not often made, perhaps because many of the samples being evaluated have L1 and L2 proficiency within the average range or higher (e.g., Bialystok, Craik & Ruocco, 2006; Lee Salvatierra & Rosselli, 2011). However, the need for such a distinction is amplified for high-risk populations such as ELs with reading difficulties, who are more likely to have language proficiency falling below expectation in one or both languages (Kieffer, 2008). For instance, identification and intervention programs may be more likely to target a balanced bilingual with low proficiency in both languages, as opposed to a balanced bilingual who is highly proficient in each language.

We are aware of only a handful of studies considering both language proficiency and balance for Spanish–English speaking bilinguals (Lonigan et al., 2018; Vaughn & Hernandez, 2018; White & Greenfield, 2017). For instance, in a sample of adults, Vaughn and Hernandez (2018) used an equation developed by David Francis (personal communication, June 20, 2019) to produce a continuous metric of language proficiency. The equation (shown below) computes an additive combination of Spanish and English language composite scores as well as a function that provides a "boost" in score for individuals who are more balanced in their language abilities. In the study from Vaughn and Hernandez (2018), language ability was measured with two objective tests in each language: picture naming and passage comprehension. The metric integrating proficiency and balance proved useful as an outcome measure in their study, as it was found to be predicted by the interaction of genetic variants associated with cognitive flexibility and learning, as well as age of second language acquisition. This method of characterizing language has not been evaluated among bilingual children or among at-risk bilinguals.

$$(L1 + L2)\sqrt{\frac{2 * L1 * L2}{L1^2 + L2^2}} \quad (1)$$

Lonigan et al. (2018) administered objective language measures in English and Spanish to a sample of preschoolers and used latent profile analysis (LPA) to identify subgroups, which were then compared on early literacy skills. They administered tests of auditory comprehension (a complex receptive vocabulary measure) and expressive vocabulary in both languages. Although nine distinct profiles of L1 and L2 proficiency and balance were noted, the researchers focused on three "super" profiles: English Language Learners, Balanced Bilinguals, and Spanish Language Learners. The Lonigan et al. (2018) study clearly informs the characterization of language proficiency and balance in at-risk ELs, but it is unclear how such groupings would emerge in a middle school sample of students who are further identified as struggling readers, or how resultant profiles might relate to language-specific external measures such as a self-report measure of language usage.

Current study

Taken together, there is significant heterogeneity in characterizing bilingual samples in terms of language proficiency and balance. We are not aware of any studies that have

systematically compared these approaches, or of any studies that use such methods to characterize language among middle school ELs who are also struggling readers. Thus, the overarching goal of this study is to evaluate measurement approaches involved in the characterization of language proficiency and balance in a sample of middle school ELs with reading difficulties. By doing so, we hope to fill important methodological gaps while simultaneously informing language characterization approaches for a high-risk and understudied population, where such knowledge may help inform identification and intervention approaches. Importantly, we also expect that evaluating such methodological approaches in this population may draw attention to limitations associated with utilizing available language measures in the L1 (English) for at-risk bilingual samples, as such students are not represented in the normative samples used to develop such measures.

We begin with an investigation of the dimensionality of language through confirmatory factor analysis (CFA) with a wide range of assessments (i.e., expressive, receptive, syntax, and semantics, in both English and Spanish). Next, we utilize LPA with the same language assessments to determine which subgroups are present within our sample and how they are characterized in terms of L1 and L2 proficiency as well as balance. It will then be possible to compare the variable-centered approach (i.e., CFA) with the person-centered approach (i.e., LPA) to evaluate their convergence. The identified latent profiles will then be compared on a closely related single metric of proficiency and balance (i.e., the equation developed by Francis that appears in Vaughn & Hernandez, 2018) as well as a self-report measure of language usage.

Hypotheses

1. Based on prior factor analytic work in bilingual samples of children (i.e., Gottardo & Mueller, 2009; LARRC et al., 2018), we predict that objective language measures will disaggregate according to language (i.e., one English factor, one Spanish factor). We will also test models differentiating between semantics/syntax and expressive/receptive skills; if further differentiation occurs, we expect it to be along the dimension of semantics/syntax. Results from the best-fitting model will be used to create proficiency factor scores in English and Spanish.
2. Given that the sample of students is at risk (ELs, struggling readers, from under-resourced schools), we expect LPA using the full battery of nine objective language measures to reflect four subgroupings of level and balance within our sample: (1) balanced average proficiency; (2) balanced low proficiency; (3) unbalanced with higher English proficiency; and (4) unbalanced with higher Spanish proficiency.
3. We expect latent profiles to clearly differ on their English and Spanish proficiency factor scores computed as above, demonstrating convergence between these approaches. Specifically, we hypothesize that our balanced/average group and unbalanced/higher English group will demonstrate higher English factor scores than the balanced/low and unbalanced/higher Spanish groups, and that our balanced/average and unbalanced/higher Spanish groups

will demonstrate higher Spanish factor scores than the unbalanced/low and unbalanced/higher English groups.

4. We predict that our latent profiles will differ on a single objective metric of proficiency and balance (Vaughn & Hernandez, 2018), such that individuals in the balanced average proficiency group will have the highest scores on this metric.

5. We expect the resultant latent profiles to differ on a self-report measure of language usage; specifically, we expect that students characterized by a profile with higher English proficiency will report a higher level of English usage relative to Spanish, and vice versa for students with a profile characterized by higher Spanish proficiency. We expect a balanced level of self-reported usage among students with a more balanced profile.

Methods

Participants

Participants were 161 6th and 7th graders from public schools in the southwestern United States who were all designated as struggling readers based on failure of the statewide standardized reading test the prior year. This sample is a randomly selected subset of a larger parent study (n = 410) that included other measures and a reading intervention, as described elsewhere (Capin, Miciak, Steinle, Hamilton, Fall, Roberts, Fletcher & Vaughn, 2022). However, the current study is focused on pretest data to mitigate any effects of intervention on the means of and/or the covariances among the language measures. In accordance with the parent project, inclusion criteria for all participants included: (1) enrolled in 6th or 7th grade; (2) identified as ELs or former ELs who have been re-designated as English proficient within the last three years based on statewide assessments of listening, speaking, reading, and writing in English (all students spoke Spanish and English); (3) a parent reported that Spanish is spoken in the home at initial school entry; (4) a parent reported that their child was of Mexican or Central American origin. The restriction of ancestry to those of Mexican or Central American descent was necessary to reduce heterogeneity of the sample for the epigenetics portion of the larger parent project. Moreover, the majority of students in the middle schools served by the researchers, as well as the local communities, reflect this demographic. Exclusionary criteria for the parent project included: (1) a sensory disorder that precluded participation in the assessment and intervention protocols; and (2) participation in an alternative curriculum (i.e., life skills course).

As noted from inclusion criteria, all students were Hispanic. Forty-eight percent of students were in 6th grade and 41% were female. The mean age of the students was 12.5 years (SD = 0.75 years). Seventeen percent of the sample had been previously identified by their school as requiring special education services, though additional information about special education designation and associated interventions was not available. Seventy-six percent of the sample was identified as qualifying for free/reduced lunch, a proxy for low socioeconomic status. There were six schools and 27 classrooms (~6 students per classroom).

Procedures

All procedures were approved by the investigators' respective Institutional Review Boards. Recruitment methods included approvals at the district and school (principal) level. Teachers in grades 6 and 7 were also briefed about the study. Informed parent permission letters were sent home to students' families, and assent was obtained from students. All assessments were administered by trained, supervised data collectors, including bilingual individuals.

Measures

Three types of measures were obtained from participants: demographic information, objective language tests, and self-reported language use. We conducted objective assessments of various language constructs and administered a self-report questionnaire evaluating language usage across a range of activities and contexts. Objective language assessment included measures (in both Spanish and English) of expressive vocabulary, receptive vocabulary, expressive syntax/grammar, and receptive syntax/grammar.

Demographics

Information regarding students' gender, age, socioeconomic status, and eligibility for special education services was obtained and reported for descriptive purposes.

Language measures

Students were given assessments of semantics (both receptive and expressive) and syntax (both receptive and expressive) in both Spanish and English. The *WJ-III Picture Vocabulary* (Woodcock, McGrew, Mather & Schrank, 2007) assesses expressive semantics. The subtest requires the student to provide a single word or phrase that matches pictured stimuli. The *Woodcock-Muñoz Batería III Picture Vocabulary* (Batería III; Muñoz-Sandoval, Woodcock, McGrew & Mather, 2007) is the equivalent task in Spanish. Psychometric properties in both English and Spanish are good, with test-retest reliabilities exceeding .85 at this age. The *Receptive One Word Picture Vocabulary Test* (ROWPVT-4; Martin & Brownell, 2011) assesses receptive semantic knowledge and evaluates a student's ability to match a spoken word with an image of an object, action, or concept. The *ROWPVT-4, Spanish/Bilingual Edition* (Martin, 2013) is a measure of BILINGUAL receptive language and thus items are administered in Spanish and/or English. Items are first presented in the language that the examiner believes to be dominant for that particular student: if correct, credit is given; but, if incorrect, the same item is re-presented in the other language. However, this study required a score that reflected Spanish receptive vocabulary only: therefore, each item was given in Spanish first. The item was administered in English only if incorrect in Spanish. Thus, the resultant standard score still reflects overall receptive vocabulary in both English and Spanish, but this method of administration also allowed for a score that reflected Spanish vocabulary only. The correlation between the standard score obtained from typical administration and our Spanish-only raw score was strong ($r = .94$). Psychometric properties for the English and bilingual editions of the *ROWPVT-4* are good, with a test-retest reliability of 0.91 across all ages. The *WJ-III Memory for Sentences* (Woodcock et al., 2007) subtest evaluates expressive syntax and requires the student to remember and repeat single words, phrases, and sentences presented orally, with increasing grammatical complexity.

The *Woodcock-Muñoz Batería III Memory for Sentences* (Muñoz-Sandoval et al., 2007) is the equivalent task in Spanish, and both English and Spanish tasks have a median reliability of .89 at this age. The *Sentence Assembly* subtest from the Clinical Evaluation of Language Fundamentals-Fourth Edition (CELF-4; Semel, Wiig, Secord & Langdon, 2006) is an additional test of expressive syntax in English and assesses a student's ability to formulate syntactically and semantically correct sentences after the visual and verbal presentation of words. The CELF-4 has demonstrated adequate psychometric properties, with Cronbach's alpha ranging from .70–.91 across subtests. The *WJ-III Understanding Directions* (Woodcock et al., 2007) subtest is a measure of receptive syntax that requires the student to listen to a sequence of instructions and follow directions by pointing to various objects in a colored picture. The *Woodcock-Muñoz Batería III Understanding Directions* (Muñoz-Sandoval et al., 2007) is the analogous task in Spanish. Psychometric properties in both English and Spanish are good, with a median reliability of .77 at this age. Reliabilities (Cronbach's alpha) for these measures in our sample were adequate and are reported in Table 1.

Self-report language measure

The *ROWPVT-4, Spanish/Bilingual Edition* contains a self-report measure of language use using a 3-point Likert-type scale, where 1= "Mostly Spanish," 2= "Half Spanish, Half English," and 3= "Mostly English." Items assess the individual's language use across a range of contexts – including which language they use to speak to parents, siblings, peers, and teachers, as well as which language they use to read, watch television, etc. These items are not normed. Items were administered to students by an examiner individually, with opportunities for clarification, or the option to have items read aloud, as needed. We computed an average score for each student such that a higher score indicated more English usage. When the items were considered continuously, reliability within the present sample was .67. This measure was also used categorically to allow for relating balanced usage to balanced proficiency, by classifying each student into one of three groups: mostly English usage, balanced usage, and mostly Spanish usage.

Analyses

Before addressing specific hypotheses, descriptive statistics, correlations, and reliabilities were computed for all nine language measures as well as the self-report measure (see Table 1). Distributions of all language measures were inspected through histograms as well as values for skewness (between −1 and +1) and kurtosis (less than 3). Non-normality was noted on three measures, *Batería-III Memory for Sentences*, *Batería-III Picture Vocabulary*, and *ROWPVT-4* due to eleven total outliers. CFA and LPA results utilizing the nine standardized language variables were conducted with and without these individuals, and the same pattern of results was obtained. Therefore, these outliers were retained in the analyses reported below.

As noted, age-based standard scores were not available for the Spanish measure of receptive semantics (*ROWPVT-4 Bilingual Edition*) given how this measure was administered. Therefore, standardized raw scores for all nine language measures were used in the factor analyses and latent profile analysis, and we report both standardized raw scores and

age-based standard scores in Table 1 (but note that the standard score reported for the *ROWPVT-4 Bilingual Edition* is the score obtained from typical test administration).

A variable-centered approach was used to test Hypothesis 1. Specifically, CFA models were tested including a unitary model and three two-factor models (i.e., along the dimensions of syntax/semantics, expressive/receptive, and English/Spanish) in Mplus (Muthén & Muthén, 2012) using maximum likelihood estimation with robust standard errors (MLR). Additionally, the type = complex option was used in Mplus across confirmatory models due to the multilevel nature of the data to obtain more accurate standard errors (Snijders & Bosker, 2011). Moreover, in order to account for possible clustering effects of students within classrooms on the factor structure of the language tests, we also ran the models with the language scores centered at the classroom means.[1]

Model fit was evaluated with the chi-square statistic as well as a combination of absolute, parsimonious, and comparative fit indices. The standardized root-mean-square residual (SRMR) was used as an index of absolute fit. SRMR values less than .08 are considered acceptable (Mueller & Hancock, 2008). The root-mean-square error of approximation (RMSEA) was used as an index of parsimonious fit. Values less than .08 generally suggest acceptable model fit (MacCallum & Austin, 2000). The 90% confidence interval and closeness of fit test were also reported for the RMSEA. We also included the comparative fit index (CFI), where values greater than .90 generally indicate good fit (Schumacker & Lomax, 2004). Chi-square differences between models were examined using the Satorra-Bentler (Satorra & Bentler, 2001) scaled (mean-adjusted) chi-square formula.

A person-centered approach was used to test Hypothesis 2. Specifically, LPA was used to evaluate whether students could be grouped according to their pattern of performance across the nine language tests. Construct validation of the latent profiles was accomplished through testing Hypotheses 3 through 5.

Using Mplus, our LPA analysis began with the estimation of a two-profile model, with subsequent models adding one profile until there was no longer an improvement in model fit. According to Nylund, Asparouhov, and Muthén (2007), the best model fit indices for LPA with continuous indicators are the Bayesian Information Criterion (BIC), the sample-size adjusted Bayesian Information Criterion (ABIC), and the bootstrapped likelihood ratio test (BLRT). BIC and ABIC provide indices of how efficiently the model predicts the data, with smaller values indicating better model fit. Kass and Raftery (1995) recommend that BIC and ABIC differences greater than 10 be used to indicate differences in model fit. The BLRT provides a significance test of the model with *k* profiles against the model with *k*-1 profiles. Model entropy and posterior probability values were also computed to

[1]When multi-level data are factor analyzed, clustering presents two potential challenges. One concerns the proper estimation of standard errors given non-independence across subjects in the same cluster. The second is the potential for clustering to distort the total-groups covariance matrix if the covariation in the cluster means differs from the within-cluster covariation in individual scores. With a large number of clusters, it is possible to fit the factor model at both the cluster and within-cluster levels simultaneously. However, when the number of clusters is small, the model cannot be estimated at the cluster level, but cluster effects on the within-cluster covariation can be controlled by centering observations within-clusters at the cluster means. In the present study, we addressed the potential effects of clustering on the covariance structure by fitting the model with and without centering and compared results across the two approaches. (See Khalaf, Santi, Kulesz, Bunta & Francis, 2019 for more information on these issues.)

evaluate each model. The model entropy statistic ranges from 0–1 and indexes classification certainty; prior studies have employed a cutoff of 0.80 (Hart et al., 2016; Lonigan et al., 2018). Finally, the average posterior probability for a given profile reflects the average probability of assignment to class *k* for people assigned to each of the *k* classes, where assignment is based on the maximum posterior probability. Recommendations for model selection were informed by Nylund et al. (2007) as well as other studies that have employed LPA for related purposes (i.e., Lonigan et al., 2018). Specifically, the preferred model should show significantly better fit as measured by BIC, ABIC, and BLRT.

Resultant latent profiles and posterior probabilities from the best-fitting model were used to address Hypotheses 3 through 5. Specifically, ANOVA was used to compare the profiles on proficiency factor scores (Hypothesis 3), a metric combining proficiency and balance (Hypothesis 4), and a chi-square test was used to evaluate whether or not latent profile membership was significantly related to membership in self-reported usage groups based on student ratings on the self-report measure of language usage (Hypothesis 5). In addition, we evaluated bivariate correlations between the factor scores and the self-report measure in order to evaluate the convergence between objective and subjective measures continuously.

Analyses were first run using the latent profiles in a deterministic manner such that students were assigned to their most likely profile based on posterior probabilities. However, such categorization does not take into account classification uncertainty, as an individual's probability of profile membership is generally less than 1.0. Thus, to account for classification uncertainty, we also utilized the multiple pseudo-random draws approach in SAS which involved conducting multiple (i.e., 1000) iterations of random number generation in which students were assigned to profiles based on their posterior probabilities (Bandeen-Roche, Miglioretti, Zeger & Rathouz, 1997). Analyses for Hypotheses 3–5 were then run 1000 times using the resulting datasets, and parameter estimates and *p*-values were averaged and compared with the initial deterministic results. For example, a student whose most likely profile membership is Profile 1 but whose posterior probabilities reflect an 85% chance of being placed into Profile 1 and a 15% chance of being placed into Profile 2 would be placed into Profile 2 in approximately 15% of the 1000 simulations. Thus, averaging results over these simulations controls for the uncertainty of profile membership (Bandeen-Roche et al., 1997). If average posterior probabilities across profiles are high, then we would expect these two sets of results to converge with one another, and to vary little from sample to sample.

Results

Examination of descriptive statistics (Table 1) revealed that, on average, performance across both English (SS = 79) and Spanish (SS = 75) objective language measures fell more than one standard deviation below normative age-based expectations. Correlations among both the raw standardized scores and among the age-based standard scores are also reported in Table 1.

Variable centered results: dimensionality of language measures (Hypothesis 1)

Results from all confirmatory models can be found in Table 2. A single (conceptually appropriate) error covariance (between the English and Spanish *Memory for Sentences* measures) was added to all models. Importantly, the same pattern of results was found regardless of whether the language variables were centered at the classroom means, suggesting that the contribution of the classroom means was not altering the relationships among the language variables. The results from the initial un-centered models are presented here, though we also report the results from the classroom mean-centered models in Table 3.

Model 1 was a unitary model of all nine objective language measures, but it had a poor fit with the data, $\chi^2(26) = 135.05$, $p < .001$. Model 2, differentiating syntax and semantics measures, also demonstrated poor fit, $\chi^2(25) = 133.43$, $p < .001$, with no difference in fit between Models 1 and 2, Satorra-Bentler χ^2 difference = 1.62, $p = .203$. Model 3, differentiating expressive and receptive measures, demonstrated a non-positive definite matrix due to a perfect correlation ($r = 1.08$, 95% CI [0.91–1.25]) between the two latent variables. Finally, Model 4, a two-factor model differentiating between English and Spanish measures, provided strong fit, $\chi^2(25) = 55.09$, $p = .001$, and substantial improvement over Model 1, Satorra-Bentler χ^2 difference = 31.33, $p < .001$.

Since the English/Spanish model provided a strong fit to the data as expected, we then considered further distinctions. Specifically, a four-factor model was tested including factors for English semantics, English syntax, Spanish semantics, and Spanish syntax (Model 5). Although this model also demonstrated strong fit to the data, $\chi^2(20) = 35.95$, $p = .016$, results demonstrated a non-positive definite matrix due to a high correlation ($r = .93$, 95% CI [0.85–1.00]) between Spanish syntax and Spanish semantics. Therefore, an additional three-factor model was run including two distinct English factors (syntax and semantics) and one Spanish factor (Model 6), which showed a strong fit to the data, $\chi^2(23) = 48.37$, $p = .002$. However, results from chi-square difference comparisons demonstrated that the three-factor model did not provide a significantly better fit compared to the two-factor English/Spanish model, Satorra-Bentler χ^2 difference = 5.89, $p = .053$. Moreover, a high correlation was noted between the English latent variables ($r = .79$, 95% CI [0.43–1.14]), and for one fit index (BIC), model fit was poorer for the three-factor model. Therefore, the two-factor English/Spanish model (Model 4) was chosen as the best-fitting and most parsimonious model, supporting Hypothesis 1. This final model is shown in Figure 1. Resultant English and Spanish factor scores were utilized in subsequent analyses as indices of English and Spanish proficiency.

Person-centered results: profiles of students based on pattern of performance on objective language measures (Hypothesis 2)

Fit statistics for latent profile models are shown in Table 4. For the two- and three-profile models, the BIC, ABIC, and BLRT indicated improved fit over the *k*–1 profiles. Although the four-profile model indicated slight improvement across fit statistics, the model was not interpretable because the best log-likelihood value could not be replicated despite increasing the number of random start values, suggesting that the model was not a good fit to the data. A final five-profile model did not demonstrate improvement across all fit indices,

there were problems with non-convergence, and the best log-likelihood value could also not be replicated. Therefore, the three-profile model was chosen as the final model, providing partial support for Hypothesis 2. Entropy values across all models were acceptable (> 0.80).

The patterns of standardized raw sample means across all nine language measures for the three-profile model are shown in Figure 2, and patterns of age-based standard scores for each of the three profiles are provided in Figure 3. Inspection of the three resultant profiles demonstrated that they were characterized by differences in both proficiency in L1 and L2 as well as balance. By evaluating both standardized raw scores and age-based standard scores, we were able to better understand the pattern of student performance both relative to one another (standardized raw scores) as well as relative to normative standards based on age (standard scores).

Twenty-five percent (n = 41) of the sample was most likely to be classified into Profile 1 based on posterior probabilities. This profile was characterized by balance between Spanish and English proficiency, with higher Spanish and English scores than the other two profiles. The average age-based standard score across all Spanish measures was 85, and the average age-based standard score across all English measures was 86. The average posterior probability for Profile 1 was 91%.

Sixty-two percent (n = 100) of the sample was most likely to be classified into Profile 2 based on posterior probabilities. This profile was characterized by a moderate degree of imbalance between Spanish and English proficiency across standardized raw scores, with Spanish scores falling somewhat higher than English scores. English scores were the lowest in this profile relative to the other profiles. This pattern was reflected to some degree in age-based standard scores, although this profile's pattern of age-based standard scores actually demonstrated balance relative to their standardized raw scores. Specifically, an average standard score of 76 was noted across Spanish tests and an average of 75 was noted across English tests. The average posterior probability for Profile 2 was 91%.

Twelve percent (n = 20) of the sample was most likely to be classified into Profile 3 based on posterior probabilities. This profile was characterized by a large degree of imbalance between Spanish and English proficiency as noted in standardized raw scores, with Spanish scores far below English scores, and English scores falling in between the English scores of the other two profiles. This pattern was also reflected across age-based standard scores for all nine objective language measures, with an average Spanish score of 46 across the four Spanish measures and an average English score of 80 across the five English measures. The average posterior probability for Profile 3 was 98%.

With the pseudo-random draws procedure across 1,000 replication samples, the results were quite similar with regard to the proportion of students placed into each profile. Specifically, the average percentage of the sample placed in Profiles 1–3 across the 1,000 replications were 28%, 59%, and 13%, respectively, as compared to 25%, 62%, and 12%, respectively, when students were assigned deterministically based on their posterior probabilities.

Convergence between variable-centered and person-centered approaches (Hypothesis 3)

In order to evaluate the convergence between the variable-centered and person-centered approaches, one-way ANOVA and Tukey-Kramer multiple comparisons tests were used to examine whether the three latent profiles differed on the English and Spanish factor scores. The profiles were expected to differ on the factor scores since the same set of nine objective language measures were used in both the CFA and LPA analyses.

When the latent profiles were used in a deterministic manner, the groups differed significantly from one another on English factor scores ($F = 68.58$, $p < .001$), with Tukey-Kramer multiple comparisons demonstrating that Profile 1 had the highest English scores relative to both Profile 2 ($p < .001$) and Profile 3 ($p < .001$) with large effects as suggested by Cohen's effect size values ($d = 2.22$ and $d = 1.44$, respectively). Profile 3 demonstrated significantly higher English scores than Profile 2 ($p = .003$) with a large effect ($d = 0.83$).

The latent profiles also differed significantly from one another on Spanish factor scores ($F = 116.93$, $p < .001$), with multiple comparisons tests demonstrating that Profile 1 had the highest Spanish scores relative to both Profile 2 ($p < .001$) and Profile 3 ($p < .001$), with large effects ($d = 0.91$ and $d = 3.25$, respectively). Profile 2 demonstrated significantly higher Spanish scores than Profile 3 ($p < .001$) with a large effect (2.62).

The same pattern of results was found when latent profile classification uncertainty was accounted for using the multiple pseudo-random draws approach. Specifically, over the 1,000 iterations, ANOVA results comparing the profiles on English proficiency had an average omnibus F of 59.83 (average $p < .001$), and all pairwise comparisons were significant, on average (average $p < .01$ for all pairwise comparisons). Results comparing profiles on Spanish proficiency were also the same (average omnibus $F = 116.81$, average $p < .001$; all pairwise comparisons average $p < .001$).

Results therefore support Hypothesis 3 that the variable-centered and person-centered methods would converge with one another.

Convergence between latent profiles and continuous metric of proficiency and balance (Hypothesis 4)

One-way ANOVA and Tukey-Kramer multiple comparisons tests were used to evaluate differences across the latent profiles on a continuous metric integrating both proficiency and balance. We note that this continuous metric is closely related to both the latent profiles and the factor scores because the equation to compute the continuous metric utilized the English and Spanish factor scores. We found support for this hypothesis when considering the latent profiles deterministically, ($F = 106.35$, $p < .001$). Profile 1 had significantly higher scores on this metric relative to both Profile 2 ($p < .001$) and Profile 3 ($p < .001$), with large effects ($d = 2.21$ and $d = 3.03$, respectively). Profile 2 also performed significantly higher than Profile 3 ($p < .001$), with a large effect ($d = 1.20$). The same pattern of results was noted when considering classification uncertainty through the multiple pseudo-random draws approach (average omnibus $F = 95.03$, average $p < .001$; all pairwise comparisons average $p < .001$).

Convergence between latent profiles and self-report measure (Hypothesis 5)

We evaluated whether the latent profiles mapped onto the categorical groups based on the self-report measure of language usage. Twenty percent ($n = 32$) of the sample was classified into a group characterized by a high level of English usage, 61% ($n = 99$) into a balanced usage group, and 19% ($n = 30$) into a high Spanish usage group. A chi-square test revealed a significant relationship between deterministic latent profile membership and usage group membership, $\chi = 19.58$, $p < .001$, and this same pattern emerged when considering classification uncertainty using the pseudo-random draws approach (average $\chi = 22.20$, average $p < .001$). However, inspection of the distribution of students in LPA and usage groups indicated that not all patterns were in expected directions. Specifically, of the 32 students who reported a high level of English usage, only 11 were from Profile 3 (characterized by English dominance), with 15 from Profile 2 (moderately unbalanced, higher Spanish) and 6 from Profile 1 (balanced). Of the 99 students who reported balanced usage, only 27 were from Profile 1, whereas the majority ($n = 63$) were from Profile 2. Patterns were closer to predictions for the Spanish usage group, such that the majority (22 out of 30) of these students were from Profile 2, and the remaining 8 were from Profile 1.

We also computed bivariate correlations between the self-report measure and the English and Spanish factor scores. English factor scores demonstrated a significant but modest correlation with the self-report measure ($r = .24$, $p = .002$), whereas Spanish factor scores demonstrated a significant, moderate negative correlation with self-report ($r = -.56$, $p < .001$), reflecting a moderate positive correlation between Spanish proficiency and Spanish language usage, as lower scores on the self-report measure indicated a higher level of Spanish usage.

Discussion

The purpose of the present study was to compare approaches to characterizing both language proficiency and balance in a sample of Spanish-speaking middle school ELs further identified as struggling readers. Our results provide important information about the pattern of L1 and L2 language performance in this understudied population, highlighting its at-risk nature. Descriptively, although low English scores were expected given the EL designation of the sample as well as being selected for reading difficulties, what was striking was that Spanish scores were lower than English skills on average, and accompanied by wide variability. Particularly surprising was the pattern of performance in Profile 3, which reflected a subgroup that could essentially be characterized as monolingual English speakers given their average Spanish skills fell below the 1st percentile relative to normative expectations. In line with hypotheses, variable-centered and person-centered approaches converged with one another, and with a continuous metric integrating proficiency and balance. A self-report measure of language usage converged with objective measures, though not to the extent hypothesized.

The structure of language among middle school ELs with reading difficulties: a variable-centered view

Our CFA results clearly support the hypothesized two-factor structure of English and Spanish language skills and extend prior factor analytic work by considering these relationships in an at-risk sample of middle school English Learners who are also struggling readers. That a unidimensional model provided poor fit to the data suggests that investigations of language processes in this context should consider both English and Spanish processes rather than utilizing performance in one language to generalize to the student's overall language skills. This conclusion is consistent with Branum-Martin, Mehta, Fletcher, Carlson, Ortiz, Carlo and Francis (2006), who argued that a joint measurement model of English and Spanish tasks is needed to evaluate language among bilingual children. This is further highlighted by the low and mostly non-significant correlations between the English and Spanish measures found in this study, as well as a low correlation between the resultant factor scores ($r = .06$). In this regard, our findings are consistent with some prior factor analytic work with bilingual samples (Gottardo, 2002; Simon-Cereijido & Gutiérrez-Clellen, 2009) but inconsistent with other bilingual studies that report high correlations between English and Spanish language factors (Castilla et al., 2009; LARRC et al., 2018; Lucero, 2015). It is possible that differences in sample characteristics across studies may explain some of the differences in findings. For instance, students in the Castilla et al. (2009) and LARRC et al. (2018) studies had Spanish skills within the average range, which is different from the low to low average and widely variable Spanish skills of our sample.

Although factor models demonstrated that English and Spanish measures clustered together into distinct factors, there was a lower than expected level of coherence among the five English language measures. In fact, the highest correlation between English measures, $r = .39$, was lower than most of the intercorrelations among the Spanish measures, which ranged from $r = .36$ to $r = .70$ (see Table 1). Indices of internal consistency values for these English measures were adequate, but lower than those reported by the test developers, and also lower on average than those of the Spanish tests. These findings could potentially reflect issues of construct validity of the English measures in this unique sample, as these tests are normed on monolingual children. This will continue to be an important issue as the proportion of the population that speaks both English and Spanish continues to increase in the United States. On the other hand, it is important to note that there may be situations where direct comparisons between bilingual and monolingual performance on the same test is useful; for instance, if the purpose of the assessment is to better understand how a bilingual student's English skills directly compare to those of her monolingual peers in order to inform intervention or instructional approaches.

In contrast, a high level of cohesion among the Spanish measures, which are normed with Spanish–English speaking bilingual children, may be construed as support for the construct validity of these tests. One possible explanation for the high correlation between the Spanish syntax and vocabulary measures is related to the fact that some aspects of Spanish syntax are more closely related to Spanish vocabulary than others (Pérez-Leroux, Castilla-Earls & Brunner, 2012). For example, vocabulary growth in Spanish impacts aspects

of expressive syntactic output including aspects of sentence complexity (e.g., utterance length and subordination rates). Given that our measure of Spanish expressive syntax required the student to repeat increasingly grammatically complex sentences, it is possible that performance was influenced by level of Spanish vocabulary in addition to syntactic knowledge. For more discussion about the relation of syntactic and semantic knowledge within and across languages in young bilingual children, see the work of Simon-Cereijido and Méndez (2018). Similarly, as noted by Bates and Goodman (1999) with regard to measurement of language processes in monolinguals, it is impossible to test an individual's grammatical knowledge without also evaluating their semantic knowledge given the strong longitudinal association between these skills in early language development. Therefore, it is possible that the syntax measures we employed in our study were dependent on semantic knowledge and thus may not have adequately captured syntactic ability.

Classifying students by proficiency and balance: a person-centered view

Our LPA results showed profiles of both language proficiency levels as well as balance between English and Spanish skills, but only three of our hypothesized four profiles were obtained. We anticipated two balanced groups (one with higher proficiency levels, one with lower proficiency levels) and two unbalanced groups (one with English skills higher than Spanish, another with Spanish skills higher than English). Inspection of standardized raw score performance across our three profiles demonstrated each of these expected categories except the balanced-lower proficiency group, though we note that, on average, our sample performed in the low average range across all tests.

Although we utilized standardized raw scores in our models, also considering the patterns of age-based standard scores across the latent profiles allowed us to understand relative levels of performance within the sample as well as relative levels compared to normative samples used to develop the assessments. While these two patterns of scores were consistent for Profiles 1 and 3, there were discrepancies between these two approaches for Profile 2 with regard to interpreting level of English proficiency. Specifically, the pattern of standardized scores for Profile 2 indicated that these students were characterized by a moderate level of imbalance between English and Spanish skills (with higher Spanish skills). In contrast, the pattern of age-based standard scores was very similar across all nine language measures for Profile 2, yielding a balanced profile, yet one that is significantly lower than that of Profile 1. We are inclined to consider the standardized raw score results, as the age-based standard scores were obtained from norms from five different normative samples (*Batería-III, CELF-4, ROWPVT-4, ROWPVT-4 Spanish/Bilingual Edition*, and *WJ-III*), though as noted earlier there may be practical reasons for using and interpreting age-based standard scores.

Importantly, our findings highlight the wide variability in Spanish skills relative to English in this sample, with students in Profile 3 performing well below age expectations, on average, across all five Spanish measures. The wider variability in Spanish relative to English skills may reflect these students' English-speaking classroom environment, with the effect of making their English use/exposure somewhat more homogeneous. In contrast, our sample may differ in the extent to which they use/are exposed to Spanish in their home and

community environments. Age of second language acquisition and age of arrival to the US are other factors that may account for this variability (Hernandez & Li, 2007). Although this data was unavailable to us, more information regarding the students' instructional history regarding language exposure, as well as history of language exposure in the home and community throughout development, would be helpful in further contextualizing our findings.

To our knowledge, this is the first study to demonstrate convergence between variable-centered and person-centered approaches to characterizing language. Comparison of such approaches has been conducted in other areas, such as academic self-concept (Marsh, Lüdtke, Trautwein & Morin, 2009) and prejudice (Meeusen, Meuleman, Abts & Bergh, 2018), and is important because these two approaches address different yet complementary questions (i.e., factor analysis addresses questions about the relationships among measures, whereas LPA addresses questions about subgroupings of individuals), and studies often choose one approach or the other rather than considering both. Our results suggest that both approaches reach a similar conclusion regarding the characterization of proficiency levels and balance, and this confidence was buoyed by our use of the multiple pseudo-random draws procedure. Our findings therefore support the use of these procedures in characterizing language. Since the extent to which our specific pattern of results (e.g., three-profile model) would hold in a different population (e.g., with different L1 and L2 proficiency levels, presence or absence of risk factors, age level, etc.) is unclear, future work should consider applying these methods with different bilingual samples in different contexts.

Convergence of objective language measures with a self-report measure

We included a self-report measure as a construct validity target to demonstrate its convergence/divergence with objective measures. Whether or not self-report measures map onto objective test results is also an important question, since self-report scales are often utilized as a proxy for language across studies. Moreover, in clinical contexts (e.g., neuropsychological evaluation, psychoeducational testing), self-reported information about language levels and language usage may be used to inform decisions regarding the language of assessment. We did find that LPA group membership (based on objective tests) was significantly associated with usage group membership (based on self-report). However, students with either balanced proficiency (Profile 1) or moderately unbalanced proficiency (Profile 2) did not differentiate into usage groups in expected directions, as most students from these profiles were reporting balanced usage. In line with hypotheses though, none of the students with low Spanish proficiency (Profile 3) reported greater Spanish than English usage.

These results suggest that the self-report measure was able to differentiate between students of different proficiency levels only when there were very large proficiency differences, as the profiles had far greater variability in their Spanish skills, particularly Profile 3 relative to the other two groups. Taken together, these findings suggest that reliance on self-report of current language use alone is unlikely to provide a full picture of a students' English and Spanish skills. It is possible that additional information regarding language usage such

as age of L2 acquisition and usage throughout development would have demonstrated a stronger relationship to objective measures.

Limitations and future directions

Findings from this study should be considered in light of a few limitations. First, the lack of English language assessments developed for and normed with bilingual children is a drawback of this research. Utilizing a set of separate norms that more closely resembles our sample would be important to shed light on our sample's language abilities within the context of other Spanish-English speaking ELs, which would likely reflect a higher level of performance than the available norms used in this study. Such an approach would likely influence our conclusions regarding level of language difficulties in our sample. However, as such assessments are currently unavailable, our study draws important attention to this issue by reporting low correlations between English measures, lower reliability values on average for English tests relative to Spanish tests, and low correlations between the self-report measure and English proficiency despite a moderate relationship between the self-report measure and Spanish proficiency. Nevertheless, as noted, there is some utility of using the current norms in that they reflect the performance of our sample relative to their peers at school, which can serve as a marker for informing the services and interventions that may be of benefit.

A second limitation is the sub-optimal reliability found for the self-report measure, which may also be related to problems with the scaling of this measure, making it difficult to draw conclusions about its utility in characterizing bilinguals alongside objective measures. Future work should develop measures designed for the specific purpose of characterizing language usage and perceived proficiency in bilingual samples of children. While such assessments are currently unavailable, one approach could be to modify instruments developed and validated for these purposes in adult bilingual samples, many of which are currently in development or have been developed recently (e.g., Anderson et al., 2018).

A final limitation may be the somewhat restricted range of language proficiency in our sample, though the purpose of our study was to investigate the characterization of language in this at-risk population, and we did find substantial heterogeneity even within this restricted range. However, in order to better understand relationships among English proficiency, Spanish proficiency, balanced proficiency, and self-report, future work should replicate the current study by considering these variables in a larger sample of English-Spanish bilinguals with a greater range of proficiency levels than was represented in this study.

Though not a limitation that could be resolved in the methodology of the current study, a systemic issue that is relevant to the generalization of these findings to ELs relates to the ways in which students are designated as ELs. Since students are repeatedly tested throughout their schooling to inform decisions about EL designation, this dynamic approach makes it difficult to draw conclusions about ELs as a population. As suggested by Saunders and Marcelletti (2013), a more static classification system that differentiates between students who have ever been designated as EL (including those with a current designation as well as those who have been reclassified as fluent English proficient) and students who

have never been designated as EL may be a more informative approach to studying this population.

Summary

Our study is the first to systematically evaluate the characterization of language proficiency and balance using both variable-centered and person-centered approaches, and both objective and self-report measures. Our findings reflect the multidimensionality of language in this important and understudied context (along the English/Spanish dimension), a three-profile group structure, convergence between variable-centered and person-centered methods, and partial support for the use of self-report tools. Future studies should consider additional tools for measuring self-report language variables in this population. Importantly, our results highlight the heterogeneity of language skills among middle school Spanish-speaking English Learners who are struggling readers, suggesting that future work should consider how this heterogeneity relates to important outcomes. Variability in proficiency and balance may have particular significance for language-related processes such as reading, and it will be important to directly test these relations as a means of evaluating the external validity of these metrics.

Acknowledgments.

This research was supported by F31 HD098797 awarded to the first author, and P50 HD052117, awarded to Jack Fletcher, Ph.D., Texas Center for Learning Disabilities, both from the Eunice Kennedy Shriver National Institute of Child Health & Human Development to the University of Houston. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Eunice Kennedy Shriver National Institute of Child Health & Human Development or the National Institutes of Health.

Data Availability Statement.

Data are available upon request from the corresponding author.

References

American Psychiatric Association (2013) Diagnostic and Statistical Manual of Mental Disorders (DSM-5®). American Psychiatric Pub.

Anderson JA, Mak L, Chahi AK, and Bialystok E (2018) The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. Behavior Research Methods 50, 250–263. [PubMed: 28281208]

Archila-Suerte P, Woods EA, Chiarello C, and Hernandez AE (2018) Neuroanatomical profiles of bilingual children. Developmental Science 21, e12654. [PubMed: 29480569]

Balfanz R, Herzog L, and Mac Iver DJ (2007) Preventing student disengagement and keeping students on the graduation path in urban middle-grades schools: Early identification and effective interventions. Educational Psychologist 42, 223–235.

Bandeen-Roche K, Miglioretti DL, Zeger SL, and Rathouz PJ (1997) Latent variable regression for multiple discrete outcomes. Journal of the American Statistical Association 92(440), 1375–1386.

Bates E, and Goodman JC (1999) On the emergence of grammar from the lexicon. In MacWhinney B (Ed.), The emergence of language. Lawrence Erlbaum Associates Publishers, pp. 29–79.

Bedore LM, Peña ED, Summers CL, Boerger KM, Resendiz MD, Greene K, Bohman TM, and Gillam RB (2012) The measure matters: Language dominance profiles across measures in Spanish–English bilingual children Bilingualism: Language and Cognition 15, 616–629. [PubMed: 23565049]

Bialystok E, Craik FI, and Ruocco AC (2006) Dual-modality monitoring in a classification task: The effects of bilingualism and ageing. Quarterly Journal of Experimental Psychology 59, 1968–1983.

Bloom L, and Lahey M (1978) Language development and language disorders John Wiley & Sons, Inc.

Bradley RH, and Corwyn RF (2002) Socioeconomic status and child development. Annual Review of Psychology 53, 371–399.

Branum-Martin L, Mehta PD, Fletcher JM, Carlson CD, Ortiz A, Carlo M, and Francis DJ (2006) Bilingual phonological awareness: Multilevel construct validation among Spanish-speaking kindergarteners in transitional bilingual education classrooms. Journal of Educational Psychology 98, 170–181.

Capin P, Miciak J, Steinle P, Hamilton B, Fall AM, Roberts G, Fletcher J, and Vaughn S (2022) The effects of an extensive intervention for bilingual Latinx students with reading difficulties in middle school. [Manuscript in preparation]. Department of Special Education, The University of Texas at Austin.

Castilla AP, Restrepo MA, and Perez-Leroux AT (2009) Individual differences and language interdependence: A study of sequential bilingual development in Spanish–English preschool children. International Journal of Bilingual Education and Bilingualism 12(5), 565–580.

Foorman BR, Koon S, Petscher Y, Mitchell A, and Truckenmiller A (2015) Examining general and specific factors in the dimensionality of oral language and reading in 4th–10th grades. Journal of Educational Psychology 107, 884–899. [PubMed: 26346839]

Francis DJ, Rivera M, Lesaux N, Kieffer M, and Rivera H (2006) Practical guidelines for the education of English language learners: Research-based recommendations for instruction and academic interventions. Portsmouth, NH: RMC Research Corporation, Center on Instruction.

Gollan TH, Weissberger GH, Runnqvist E, Montoya RI, and Cera CM (2012) Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary norms for young and aging Spanish–English bilinguals. Bilingualism: Language and Cognition 15, 594–615. [PubMed: 25364296]

Goodman EBJC (1997) On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. Language and Cognitive Processes 12, 507–584.

Gottardo A (2002) The relationship between language and reading skills in bilingual Spanish-English speakers. Topics in language disorders 22(5), 46–70.

Gottardo A, and Mueller J (2009) Are first-and second-language factors related in predicting second-language reading comprehension? A study of Spanish-speaking children acquiring English as a second language from first to second grade. Journal of Educational Psychology 101, 330–344.

Halle T, Hair E, Wandner L, McNamara M, and Chien N (2012) Predictors and outcomes of early versus later English language proficiency among English language learners. Early Childhood Research Quarterly 27, 1–20. [PubMed: 22389551]

Hammer CS, Jia G, and Uchikoshi Y (2011) Language and literacy development of dual language learners growing up in the United States: A call for research. Child Development Perspectives 5, 4–9. [PubMed: 23259006]

Hart SA, Logan JA, Thompson L, Kovas Y, McLoughlin G, and Petrill SA (2016) A latent profile analysis of math achievement, numerosity, and math anxiety in twins. Journal of educational psychology 108, 181–193. [PubMed: 26957650]

Hernandez AE, and Li P (2007) Age of acquisition: its neural and computational mechanisms. Psychological Bulletin 133, 638. [PubMed: 17592959]

Hoff E (2013) Interpreting the early language trajectories of children from low-SES and language minority homes: implications for closing achievement gaps. Developmental Psychology 49, 4–14. [PubMed: 22329382]

Kass RE, and Raftery AE (1995) Bayes factors. Journal of the American Statistical Association 90, 773–795.

Kieffer MJ (2008) Catching up or falling behind? Initial English proficiency, concentrated poverty, and the reading growth of language minority learners in the United States. Journal of Educational Psychology 100(4), 851.

Khalaf S, Santi KL, Kulesz PA, Bunta F, and Francis DJ (2019) Bilingual phonological awareness: Construct validation in Grade 1 Spanish-speaking English learners. New Directions for Child and Adolescent Development 2019, 79–110. [PubMed: 31264340]

Kim DH, Lambert RG, and Burts DC (2018) Are young dual language learners homogeneous? Identifying subgroups using latent class analysis. The Journal of Educational Research 111, 43–57.

LARRC (Language and Reading Research Consortium), Yeomans-Maldonado G, Bengochea A, and Mesa C (2018) The dimensionality of oral language in kindergarten Spanish–English dual language learners. Journal of Speech, Language, and Hearing Research 61, 2779–2795. 10.1044/2018_JSLHR-L-17-0320

Lee Salvatierra J, and Rosselli M (2011) The effect of bilingualism and age on inhibitory control. International Journal of Bilingualism 15, 26–37.

Li P, Sepanski S, and Zhao X (2006) Language history questionnaire: A web-based interface for bilingual research. Behavior research methods 38, 202–210. [PubMed: 16956095]

Li P, Zhang FAN, Tsai E, and Puls B (2014) Language history questionnaire (LHQ 2.0): A new dynamic web-based research tool. Bilingualism: Language and Cognition 17(3), 673–680.

Lonigan CJ, Goodrich JM, and Farver JM (2018) Identifying differences in early literacy skills across subgroups of language-minority children: A latent profile analysis. Developmental psychology 54, 631–647. [PubMed: 29251963]

Lonigan CJ, and Milburn TF (2017) Identifying the dimensionality of oral language skills of children with typical development in preschool through fifth grade. Journal of Speech, Language, and Hearing Research 60, 2185–2198.

Lucero A (2015) Cross-linguistic lexical, grammatical, and discourse performance on oral narrative retells among young Spanish speakers. Child Development 86(5), 1419–1433. [PubMed: 26082153]

MacCallum RC, and Austin JT (2000) Applications of structural equation modeling in psychological research. Annual review of psychology 51, 201–226.

MacWhinney B (2008) A unified model. Routledge/Taylor & Francis Group.

Marian V, Blumenfeld HK, and Kaushanskaya M (2007) The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. Journal of Speech, Language, and Hearing Research 50, 940–967.

Marsh HW, Lüdtke O, Trautwein U, and Morin AJ (2009) Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. Structural Equation Modeling: A Multidisciplinary Journal 16, 191–225.

Martin NA (2013) Receptive one-word picture vocabulary test: Spanish-bilingual edition (4th ed.). Novato, CA: Therapy Publications.

Martin NA, and Brownell R (2011) Receptive one-word picture vocabulary test (4th ed.). Austin, TX: Pro-Ed, Inc.

Meeusen C, Meuleman B, Abts K, and Bergh R (2018) Comparing a variable-centered and a person-centered approach to the structure of prejudice. Social Psychological and Personality Science 9, 645–655.

Mueller RO, and Hancock GR (2008) Best practices in structural equation modeling. Best practices in quantitative methods 488–508.

Muñoz-Sandoval AF, Woodcock RW, McGrew KS, and Mather N (2007) Batería III: Woodcock-Muñoz: Pruebas de Aprovechamiento. Itasca, IL: Riverside Publishing.

Muthén LK, and Muthén BO (2012) Mplus statistical modeling software: Release 7.0. Los Angeles, CA: Muthén & Muthén.

National Center for Educational Statistics (2003) National Assessment of Educational Progress, 2003, Reading Assessments. Washington, DC: U.S. Department of Education, Institute of Education Sciences.

Nylund KL, Asparouhov T, and Muthén BO (2007) Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. Structural equation modeling: A multidisciplinary Journal 14, 535–569.

Ortman JM, and Shin HB (2011, August). Language projections: 2010 to 2020. In American Sociological Association Annual Meeting.

Passel JS, Cohn DV, and Lopez MH (2011) Hispanics account for more than half of nation's growth in past decade. Washington, DC: Pew Hispanic Center, 1–7.

Pérez-Leroux AT, Castilla-Earls AP, and Brunner J (2012) General and specific effects of lexicon in grammar: Determiner and object pronoun omissions in child Spanish. Journal of Speech, Language, and Hearing Research 55, 313–327.

Pinker S (1998) Words and rules. Lingua 106, 219–242.

Rosselli M, Ardila A, Lalwani LN, and Vélez-Uribe I (2016) The effect of language proficiency on executive functions in balanced and unbalanced Spanish–English bilinguals. Bilingualism: Language and Cognition 19, 489–503.

Sandhofer C, and Uchikoshi Y (2013) Cognitive consequences of dual language learning: Cognitive function, language and literacy, science and mathematics, and social-emotional development. Espinosa L (Ed.), California's best practices for teaching young dual language learners: Research overview papers. Sacramento, CA: California Department of Education.

Satorra A, & Bentler PM (2001) A scaled difference chi-square test statistic for moment structure analysis. Psychometrika 66(4), 507–514.

Saunders WM, and Marcelletti DJ (2013) The gap that can't go away: The catch-22 of reclassification in monitoring the progress of English learners. Educational Evaluation and Policy Analysis 35, 139–156.

Schumacker RE, and Lomax RG (2004)A beginner's guide to structural equation modeling. Psychology Press.

Semel EM, Wiig EH, Secord W, and Langdon HW (2006)CELF 4: Clinical Evaluation of Language Fundamentals 4: Spanish Edition. PsychCorp.

Sheng L, Lu Y, and Gollan TH (2014) Assessing language dominance in Mandarin–English bilinguals: Convergence and divergence between subjective and objective measures. Bilingualism: Language and Cognition 17, 364–383. [PubMed: 25379011]

Simon-Cereijido G, and Gutiérrez-Clellen VF (2009) A cross-linguistic and bilingual evaluation of the interdependence between lexical and grammatical domains. Applied Psycholinguistics 30, 315–337. [PubMed: 19444336]

Simon-Cereijido G, and Méndez LI (2018) Using language-specific and bilingual measures to explore lexical–grammatical links in young latino dual-language learners. Language, Speech, and Hearing Services in Schools 49, 537–550. [PubMed: 29625426]

Snijders TA, and Bosker RJ (2011) Multilevel analysis: An introduction to basic and advanced multilevel modeling. Sage.

Tomblin JB, and Zhang X (2006) The dimensionality of language ability in school-age children. Journal of Speech, Language, and Hearing Research 49, 1193–1208.

Vaughn KA, and Hernandez AE (2018) Becoming a balanced, proficient bilingual: Predictions from age of acquisition & genetic background. Journal of neurolinguistics 46, 69–77. [PubMed: 30038460]

Vega C, and Fernandez M (2011) Errors on the WCST correlate with language proficiency scores in Spanish–English bilingual children. Archives of Clinical Neuropsychology 26, 158–164. [PubMed: 21148172]

White LJ, and Greenfield DB (2017) Executive functioning in Spanish-and English-speaking Head Start preschoolers. Developmental Science 20, e12502.

Wiig EH, Secord WA, and Semel E (2013)Clinical evaluation of language fundamentals: CELF-5. Pearson.

Woodcock RW, McGrew KS, Mather N, and Schrank FA (2007) Woodcock-Johnson normative update tests of cognitive abilities. Rolling Meadows, IL: Riverside.

Yow WQ, and Li X (2015) Balanced bilingualism and early age of second language acquisition as the underlying mechanisms of a bilingual executive control advantage: why variations in bilingual experiences matter. Frontiers in Psychology 6, 164. [PubMed: 25767451]

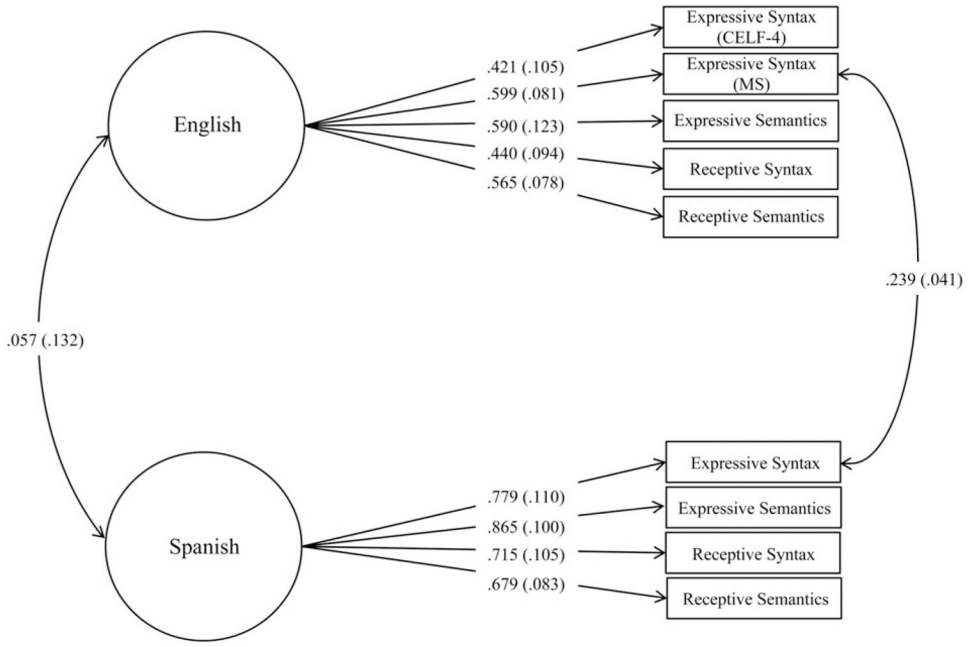

Figure 1.
Two-Factor Model

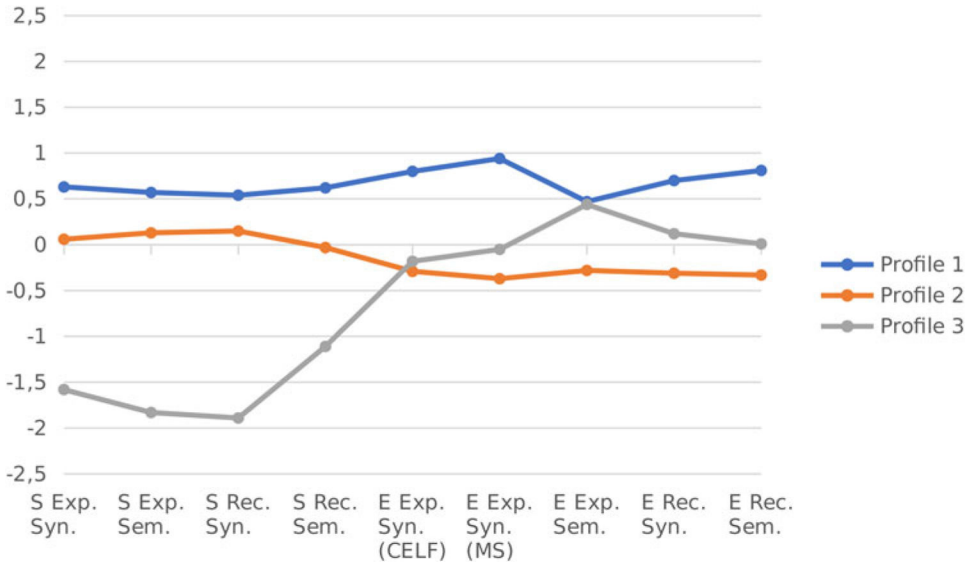

Figure 2.

Standardized Raw Score Performance on Language Measures across Latent Profiles

Note. S Exp. Syn. = Spanish Expressive Syntax; S Exp. Sem. = Spanish Expressive Semantics; S Rec. Syn. = Spanish Receptive Syntax; S Rec. Sem. = Spanish Receptive Semantics; E Exp. Syn. (CELF) = English Expressive Syntax measured with the Clinical Evaluation of Language Fundamentals Sentence Assembly subtest; E Exp. Syn. (MS) = English Expressive Syntax measured with the Woodcock Johnson – Third Edition Memory for Sentences subtest; E Exp. Sem. = English Expressive Semantics; E Rec. Syn. = English Receptive Syntax; E Rec. Sem. = English Receptive Semantics.

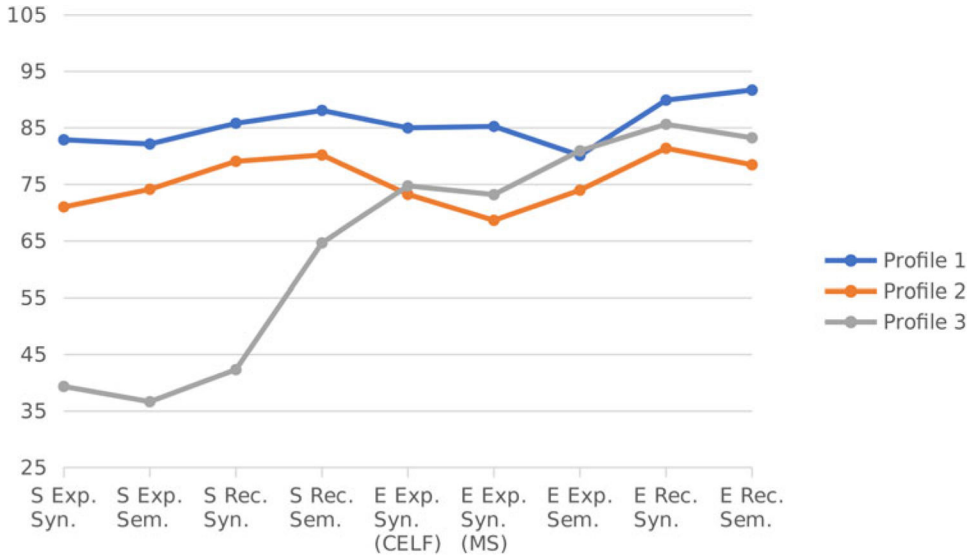

Figure 3.

Age-Based Standard Score Performance on Language Measures across Latent Profiles

Note. S Exp. Syn. = Spanish Expressive Syntax; S Exp. Sem. = Spanish Expressive Semantics; S Rec. Syn. = Spanish Receptive Syntax; S Rec. Sem. = Spanish Receptive Semantics; E Exp. Syn. (CELF) = English Expressive Syntax measured with the Clinical Evaluation of Language Fundamentals Sentence Assembly subtest; E Exp. Syn. (MS) = English Expressive Syntax measured with the Woodcock Johnson – Third Edition Memory for Sentences subtest; E Exp. Sem. = English Expressive Semantics; E Rec. Syn. = English Receptive Syntax; E Rec. Sem. = English Receptive Semantics.

Table 1.

Correlations among Language Measures, Descriptive Statistics, and Reliabilities

		1	2	3	4	5	6	7	8	9	10
1	E WJ-III Memory for Sentences	--	.34**	.34**	.29**	.27**	.30**	.07	.03	.13	.12
2	E CELF-4 Sentence Assembly	.34**	--	.15	.23*	.24*	.15	.09	.09	.15	.07
3	E WJ-III Picture Vocabulary	.34**	.18*	--	.23*	.38**	−.20*	−.14	−.16*	−.02	.39**
4	E WJ-III Understanding Directions	.28**	.22*	.23*	--	.29**	.14	−.02	.14	.10	.01
5	E ROWPVT-4	.26**	.25*	.39**	.28**	--	.05	.08	−.02	.19*	.12
6	S Batería-III Memory for Sentences	.31**	.18*	−.20*	.13	0.05	--	.69**	.59**	.55**	−.44*
7	S Batería-III Picture Vocabulary	.06	.07	−.15	−.02	.07	.70**	--	.63**	.61**	−.59**
8	S Batería-III Understanding Directions	.04	.04	−.15	.11	−.04	.60**	.65**	--	.43**	−.44**
9	S ROWPVT-Bilingual+	.08	.18*	−.02	.12	.20*	.49**	.51**	.36**	--	−.28**
10	ROWPVT Self-Report	.12	.08	.38**	.01	.12	−.46**	−.59**	−.41**	−.20*	--
	Mean Age-Based Standard Score	73.44	5.29	76.42	84.11	82.44	70.11	71.52	76.24	80.29	2.17++
	SD	13.12	2.38	9.50	9.18	12.68	19.48	19.37	18.23	16.58	0.35
	Reliability (Cronbach's Alpha)	.71	.86	.76	.75	.95	.81	.88	.95	.98	.67

Note. Values below the diagonal represent correlations between age-based standard scores. Values above the diagonal indicate correlations between standardized raw scores. WJ-III = *Woodcock Johnson Tests of Cognitive Abilities, Third Edition*; CELF-4 = *Clinical Evaluation of Language Fundamentals, Fourth Edition*; ROWPVT-4 = *Receptive One Word Picture Vocabulary Test, Fourth Edition*; ROWPVT-Bilingual = *Receptive One Word Picture Vocabulary Test, Bilingual Edition*. All age-based standard scores reported have a mean of 100 and standard deviation of 15, with the exception of the CELF-4 which uses scaled scores with a mean of 10 and standard deviation of 3.

* $p < .05$

** $p < .001$

+ Age-based standard scores were obtained through standard administration of the test, whereas standardized raw scores were obtained through modified administration.

++ Age-based standard scores are not available for this measure. Mean and standard deviation of raw scores are reported.

Table 2.

Results from Confirmatory Factor Models

Model #	Model	LL	χ^2	*df*	*p*	MLR Scaling Factor	RMSEA, *p* close [90% CI]	CFI	SRMR	AIC	BIC
1	Unidimensional: L	−1896.52	135.05	26	<.001	1.0434	0.161, $p < .001$ [.135–.189]	0.74	0.13	3849.03	3935.31
2	2 factors: Syn + Sem	−1895.67	133.43	25	<.001	1.0434	0.164, $p < .001$ [.137–.192]	0.75	0.13	3849.34	3938.70
3	2 factors: Exp + Rec	−1895.40	138.36	25	<.001	1.0023	0.168, $p < .001$ [.141–.196]	0.73	0.13	3848.80	3938.16
4	2 factors: S + E	−1852.88	55.09	25	.001	0.9737	0.086, $p = .029$ [.055–.117]	0.93	0.07	3763.76	3853.12
5	4 factors: SSx + SSm + ESx + ESm	−1842.86	35.95	20	.016	0.9346	0.070, $p = .171$ [.030–.107]	0.96	0.06	3753.72	3858.49
6	3 factors: S + ESx + ESm	−1848.60	48.37	23	.002	0.9322	0.083, $p = .051$ [.050–.115]	0.94	0.06	3759.20	3854.73

Note. LL = log likelihood; RMSEA = root-mean-square error of approximation; CI = confidence interval; CFI = comparative fit index; SRMR = standardized root-mean-square residual; AIC = Akaike's information criteria; L = language general factor with all indicators; Syn = Syntax; Sem = Semantics; Exp = Expressive; Rec = Receptive; S = Spanish; E = English; ESx = English syntax; ESm = English semantics; SSx = Spanish syntax; SSm = Spanish semantics.

Table 3.

Results from Confirmatory Factor Models with Mean-Centered Language Variables

Model #	Model	LL	χ^2	*df*	*p*	MLR Scaling Factor	RMSEA, *p* close [90% CI]	CFI	SRMR	AIC	BIC
1	Unidimensional: L	−1738.06	114.73	26	<.001	1.180	0.146, $p < .001$ [.119–.173]	0.67	0.13	3532.13	3618.41
2	2 factors: Syn + Sem	−1735.84	107.11	25	<.001	1.222	0.143, $p < .001$ [.116–.171]	0.69	0.13	3529.68	3619.04
3	2 factors: Exp + Rec	−1736.93	11a5.71	25	<.001	1.150	0.150, $p < .001$ [.123–.178]	0.66	0.13	3531.87	3621.23
4	2 factors: S + E	−1696.01	44.07	25	.011	1.163	0.069, $p = .168$ [.033–.102]	0.93	0.07	3450.02	3539.38
5	4 factors: SSx + SSm + ESx + ESm	−1687.98	33.99	20	.026	1.036	0.066, p = .226 [.023–.103]	0.95	0.06	3443.96	3548.73
6	3 factors: S + ESx + ESm	−1694.72	46.37	23	.003	1.050	0.079, $p = .072$ [.046–.112]	0.91	0.06	3451.44	3546.96

Note. LL = log likelihood; RMSEA = root-mean-square error of approximation; CI = confidence interval; CFI = comparative fit index; SRMR = standardized root-mean-square residual; AIC = Akaike's information criteria; L = language general factor with all indicators; Syn = Syntax; Sem = Semantics; Exp = Expressive; Rec = Receptive; S = Spanish; E = English; ESx = English syntax; ESm = English semantics; SSx = Spanish syntax; SSm = Spanish semantics.

Table 4.

Results from Latent Profile Analyses

	BIC	Δ BIC	ABIC	Δ ABIC	Entropy	BLRT
2 Profiles	4029.24	165.28	3940.60	196.94	0.938	216.09***
3 Profiles	3987.32	41.92	3867.03	73.57	0.823	92.73***
4 Profiles	3971.05	16.27	3819.10	47.93	0.842	67.09***
5 Profiles	3980.05	9.00	3796.44	22.66	0.877	41.82***

Note. BIC = Bayesian Information Criterion; ABIC = Sample Size Adjusted Bayesian Information Criterion; BLRT = Bootstrapped Likelihood Ratio Test. Value reported for BLRT is two times the log likelihood difference between the two models being compared. Because the best log-likelihood value could not be replicated in the 4-profile model, the 3-profile model was chosen as the final model.

*** $p < .001$