CHARACTERIZATION OF PROFICIENCY LEVELS AND BALANCE AMONG MIDDLE SCHOOL ENGLISH LEARNERS WITH READING DIFFICULTIES: RELATIONS TO READING OUTCOMES

By

Kelly Teresa Macdonald

A Dissertation Submitted to the Department of Psychology

College of Liberal Arts and Social Sciences

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

In Clinical Psychology

Chair of Committee: Paul T. Cirino, Ph.D.

Committee Member: David J. Francis, Ph.D.

Committee Member: Arturo E. Hernandez, Ph.D.

Committee Member: Anny P. Castilla-Earls, Ph.D.

University of Houston

December 2020

DEDICATION

I dedicate my dissertation work to my supportive family and friends, especially my husband Dan and my parents, Jeff and Rhonda. I am particularly grateful to my mother, who unfortunately passed before I began my doctoral studies. Mom, thank you for instilling the values of education, perseverance, and the importance of helping others.

ACKNOWLEDGEMENTS

I would like to thank my advisor and Chair of my dissertation committee, Dr. Paul Cirino, for his invaluable expertise, encouragement, and mentorship throughout my doctoral training. I am also grateful to my committee members, Dr. David Francis, Dr. Anny Castilla-Earls, and Dr. Arturo Hernandez, for their knowledge, insight, and support throughout this project. I would also like to thank the Texas Center for Learning Disabilities research team, particularly our research coordinator, Francis Leal. I would also like to acknowledge the Eunice Kennedy Shriver National Institute of Child Health and Human Development at the National Institutes of Health for providing the funding to support this important work. I am also very grateful for the many students, parents, and teachers within the Houston Independent School District for their participation and engagement in the study. Finally, I want to thank the many friends, lab-mates, professors, and clinical supervisors who have guided me along the way throughout my doctoral studies within the Developmental Neuropsychology Lab in the Clinical Psychology program at the University of Houston.

ABSTRACT

As the linguistic diversity of the United States and the world continues to increase, the impact of bilingualism on important outcomes has been an area of interest within the domains of psychology, neuroscience, and education. Although many studies have considered group differences between bilinguals and monolinguals, the importance of examining individual differences is an emergent area of study. Specific variables of interest include language proficiency as well as balance between languages, as both have shown promise in explaining variability for outcomes such as executive function. Identifying predictors of key cognitive outcomes among bilingual samples is particularly important among at-risk groups of children and has the potential to inform intervention efforts. Although executive function is commonly studied, reading is also a promising and perhaps more important outcome to study because it is heavily rooted in language and is amenable to intervention. However, evaluating the ways in which individual variability in language proficiency and balance impacts reading in an at-risk, developmental context presupposes a strong framework by which to characterize these processes. There is currently no gold standard through which bilinguals can be characterized in terms of proficiency and balance, particularly among at-risk, younger samples.

Therefore, the overarching aims of this project are twofold: 1) to compare approaches for the characterization of proficiency and balance among an at-risk sample of children (English Learners in middle school who are further identified as struggling readers); and 2) to use these to evaluate the roles of language proficiency and balance in various reading outcomes. This work will inform theoretical and empirical gaps within the bilingual and reading literatures and holds promise for informing intervention in this at-risk context. The following

chapters therefore provide literature review, hypotheses, and methods for each aim separately (i.e., characterization of bilinguals in Chapter 1; relations to reading in Chapter 2). As noted in Chapter 2, there are aspects of the second aim that were informed by findings from Chapter 1.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1: Characterization of Proficiency Levels and Balance among Middle School English Learners with Reading Difficulties

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 important variability and individual differences, and many researchers have called for further examination of such variability (Anderson, Mak, Chahi, & Bialysok, 2018; De Feyter & Winsler, 2009; Hernandez, Denton, & Blanchard, 2011). Two important sources of variability are: 1) language *proficiency levels* in both the first (L1) and second (L2) languages; and 2) whether bilinguals are *balanced* in their L1 and L2 skills. These two factors likely have important implications for various cognitive and academic outcomes (Kim, Lambert, & Burts, 2018; Sandhofer & Uchikoshi, 2013). It is particularly important to consider these sources of bilingual variability among populations that are at risk for adverse outcomes, as such knowledge may inform identification and intervention approaches for those needing it most. However, the bilingual literature is inconsistent with regard to the methods for characterizing language proficiency levels and balance, and few studies have done so in samples of at-risk children such as English Learners (ELs). Thus, the purpose of the first part of this study is to compare various approaches to the characterization of language proficiency levels and balance in at-risk bilinguals.

Bilingualism in an English Learner, Struggling Reader Context

There may be benefits to knowing multiple languages. For instance, the "bilingual advantage" hypothesis posits that bilingual speakers in general have greater executive control, switching, and cognitive flexibility than monolinguals (Carlson & Meltzoff, 2008; Bialystok and Martin, 2004; Bialystok, Craig, & Luk, 2008; Morales, Gómez-Ariza, & Bajo, 2013). The extent to which bilingualism may confer a cognitive advantage in *at-risk* samples is not well understood, although available evidence suggests that speaking multiple languages may still provide a benefit (White & Greenfield, 2016). One particularly at-risk group is students classified as English Learners (ELs), who 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, Lesaux, Kieffer, & Rivera, 2006). 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). However, research that evaluates the measurement and characterization of language proficiency and balance among ELs, particularly at the middle school level, is limited.

Given that language skills have widespread implications for long-term outcomes (e.g., educational attainment, employment status), it is among high-risk populations such as ELs with reading difficulties where more work is urgently needed in order to better understand how L1 and L2 proficiency levels and balance may relate to important outcomes.

For instance, it is possible that variability in proficiency and balance may have implications for investigations of the "bilingual advantage" (Sandhofer & Uchikoshi, 2013). Furthermore, while the role of language proficiency in reading is widely understood (Geva & Massey-Garrison, 2013; Gough & Tunmer, 1986; Proctor, Carlo, August, & Snow, 2005), it is also possible that balance as regards bilingualism also holds relevance for reading (Grimm, Solari, Gerber, Nylund-Gibson, & Swanson, 2019). However, evaluating how L1 and L2 proficiency levels and balance impact eventual outcomes presupposes a strong measurement approach for characterizing both proficiency level, as well as balance between L1 and L2. While some existing studies have characterized both of these, the approaches used to do so are quite variable. Further, most such studies are with adults rather than in the context of at-risk youth. Given the lack of knowledge in this regard, examination of these features in an at-risk population such as middle school ELs with reading difficulties is warranted. A deeper knowledge of these relationships may have implications for the identification of academic difficulties and for interventions with this population.

Dimensionality and Measurement of Language

When considering the measurement of L1 and L2 skills through objective tests, it is important to understand how language measures relate to, versus differentiate from, one another. Language is often conceptualized as a multidimensional construct (Bloom & Lahey, 1978; Pinker, 1998) with distinctions made between type of language demands (i.e., expressive vs. receptive language) and specific language skills (i.e., semantic knowledge vs. syntactics). For example, clinical and diagnostic assessment of language often involves making a distinction between expressive and receptive abilities (DSM-5; American Psychiatric Association, 2013; CELF-5, Wiig, Secord, & Semel, 2013). Other conceptualizations distinguish between semantics (i.e., word meaning and knowledge) and syntax (i.e., grammar knowledge). For instance, Bloom and Lahey (1978) argued that language is composed of two dimensions, one of form/structure (syntax) and one of content (semantics). Similarly, Pinker (1998) proposed two distinct language systems: a mental dictionary of words and their meanings (semantics) and a set of rules for creating novel forms (syntax). An alternative to these dimensional models is a unified view of language; for example, Goodman (1997) posited that the systems of semantic and syntactic knowledge fall under a unified umbrella of lexical ability.

Empirical evidence provides support for a unified view of language in younger children, with increasing multidimensionality of semantics versus syntax as children get older. For the expressive versus receptive distinction, despite both clinical and functional relevance (Law, Garrett, & Nye, 2003), there is only weak support for the empirical distinction between these skills. A seminal study from Tomblin and Zhang (2006) utilized both exploratory and confirmatory factor analysis to examine the structure of language in a large sample of students in kindergarten, 2nd, 4th, and 8th grade. Using expressive, receptive, semantic, and syntactic measures, exploratory models revealed some differentiation between semantics and syntax, regardless of whether the measures were expressive or receptive. Thus, in their confirmatory models, Tomblin and Zhang (2006) tested unitary and two-factor semantic/syntax models across each grade level. They found that the two-factor semantics/syntax models provided only slight, non-significant improvement over the unitary models with an exception at 8th grade, where the two-factor model had a significantly better fit. Some further evidence for increasing multidimensionality comes from Lonigan and Milburn (2017) as well as Foorman et al. (2015). These studies, however, were conducted with samples of predominantly monolingual students and did not consider the structure of language among bilinguals.

Available bilingual studies evaluating the structure of language in children, while limited, also support a multidimensional view but primarily between English versus Spanish skills (Gottardo & Mueller, 2009; Gray et al., 2018). For example, Gottardo and Mueller (2009) evaluated a measurement model of language and word reading skills in both Spanish and English in a sample of 1st and 2nd grade ELs. They found that the best fitting model included two oral language factors, one English and one Spanish, each composed of measures of both semantic and syntactic abilities. In a sample of Spanish-English speaking kindergarteners, Gray et al. (2018) evaluated the dimensionality of language and found that two models provided strong fits to the data. The first was a bifactor model with a single general language factor that was explained by two second-order language factors, one English and one Spanish. The second model was a four-factor model that included one English factor and three separate Spanish factors (Spanish vocabulary, Spanish grammar, and a higher-level Spanish language factor). Of note, these studies did not take into account the clustering of students within classrooms, which is important because the covariance between L1 and L2 skills at the student level may differ from the covariance between skills at the classroom level, particularly in the younger grades when there is greater diversity among classrooms with regard to the language of instruction (Branum-Martin et al., 2009). This work suggests that among Spanish-English speaking bilinguals, language skills clearly separate between English and Spanish, but further distinctions are unclear, and their presentation at the middle school level is also unclear. 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 in order to inform measurement and characterization of language proficiency and balance.

Despite the multidimensional view of language described above, bilingual studies do not typically employ extensive language assessment. Picture vocabulary tests, which index semantic knowledge (and can be either expressive or receptive), are often used in both languages in order to determine proficiency (Gollan et al., 2012; Sheng, Lu, & Gollan, 2014; Archila-Suerte, Woods, Chiarello, & Hernandez, 2016). However, we were unable to identify any studies *comparing* the use of picture vocabulary tests to other language measures (i.e., syntactic knowledge) for the purpose of classifying L1 and L2 proficiency levels as well as balance. It is possible that investigations of the dimensionality of language could inform the selection of measures to utilize for the characterization of proficiency and balance. For example, if a two-factor structure emerged consisting of one English and one Spanish factor, then such factor scores could be used to characterize proficiency. Furthermore, in the interest of parsimony, consideration of which measures provided the strongest factor loadings on each factor may help inform which measures to use for the purposes of characterizing language. Similarly, if a four-factor structure emerged consisting of English semantics, English syntax, Spanish semantics, and Spanish syntax, for example, then this might suggest that measures assessing those four constructs provides the most robust characterization of language. However, we are not aware of any study that has evaluated this empirically, especially in an at-risk sample of children, and thus more theory-driven empirical work is needed to test these ideas.

Approaches to Measuring Language Proficiency and Balance in Bilinguals

In addition to considering the specific language processes (i.e., semantics, syntax) that may be used to characterize bilinguals, considering levels of L1 and L2 proficiency and balance is also relevant. For example, balance between L1 and L2 has been implicated in EF (e.g., Rosselli, Ardila, Lalwani, & Velez-Uribe, 2016; Vega & Fernandez, 2011; Yow & Li, 2015). However, the approaches used to assess proficiency and balance are quite variable. 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 address younger, at-risk samples, and few utilize/compare multiple approaches. In reviewing this literature, studies show variability in characterizing bilinguals along five major dimensions: (1) variability with regard to use of self-report or objective metrics; (2) variability with regard to specific *types* of measures used; (3) variability in the approach used to define *balance*; (4) 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 et al., 2016; Lonigan et al., 2018; Rosselli et al., 2016; Vaughn & Hernandez, 2018; Vega & Fernandez, 2011), self-report measures (i.e., Anderson, Mak, Chahi, & Bialystok, 2018; Kim et al., 2018; Li, Sepanski, & Zhao, 2006; Li, Zhang, Tsai, & Puls, 2014; Marian, Blumenfeld, & Kaushanskaya, 2007; Yow & Li, 2015), or a combination of the two (i.e., Gollan et al., 2012; Sheng, Lu, & Gollan, 2014; Tomoschuk, Ferreira, & Gollan, 2019). Moreover, while some studies do not include objective measures in their *characterization* of language, they do use objective measures to *validate* their self-report measures (i.e., Anderson et al., 2018; Marian et al., 2007). Across studies that consider both approaches, correlations are moderate and generally range from $r = .40$ to $r = .60$. For instance, Anderson et al. (2018) recently developed a self-report measure of language use and language proficiency that demonstrated adequate reliability and validity among a diverse sample of young adult bilinguals. Self-reported English usage correlated moderately with English *Picture Vocabulary* scores ($r = .49$). Similarly, Gollan et al. (2012) reported correlations ranging from $r = .59$ to $r = .62$ between self-reported language proficiency and picture naming performance across adult samples. When compared to objective assessment, key arguments for the use of self-report methods are that they are less time consuming, require fewer resources to administer, provide more contextual information, and can be easily adapted for use across a wide range of languages. Meta-analytic findings across disciplines report that individuals' self-report of their own abilities tends to correlate imperfectly with objective data ($r = .29$; Zell & Krizan, 2014). In the language area, the somewhat larger correlations suggest that self-report and objective measures may be either convergent or complementary, though this is infrequently examined. Thus, the present study includes both types of measures.

A second, related issue across the bilingual literature in terms of characterizing proficiency and balance is the 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 (Marian et al., 2007; Gollan et al., 2012; Sheng et al., 2014; Tomoschuk et al., 2019); still others consider both (i.e., Anderson et al., 2018; Li et al., 2006; 2014; Yow & Li, 2015). For example, a measure of language usage may ask the individual to rate the extent to which they use each language across contexts such as speaking to family members, watching television, reading, etc. In contrast, a measure assessing perceived language proficiency may ask the individual to rate their proficiency in speaking each language on a scale from 0-10. These two types of self-ratings are related to one another at a moderate level; for instance, Yow and Li (2015) reported a correlation between ratings of proficiency and ratings of usage ($r = .56$). There is also variability with regard to specific types of objective tests used across studies. Specifically, some studies use a single test across languages (i.e., picture naming; Gollan et al., 2012; Sheng et al., 2014; Tomoschuk et al., 2019), whereas others use multiple tests and compute composite scores (i.e., Archila-Suerte et al., 2016; Rosselli et al., 2016; Vaugh & Hernandez, 2018; Vega & Fernandez, 2011). The present study considers a wide range of objective language assessments in English and Spanish to contrast the roles these measures play in characterizing proficiency and balance. Moreover, objective measures will be considered alongside a self-report metric 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 use categorical approaches such as latent profiles (i.e., Kim et al., 2018; Lonigan et al., 2018), median or mean splits (i.e., Archila-Suerte et al., 2016; Rosselli et al., 2016) or other cut-off scores (i.e., Vega & Fernandez, 2011). Still other studies obtain continuous scores (i.e., Sheng et al., 2014 used a ratio of L1 to L2 such that a score of 1.0 would denote perfect balance) and then use cut-points to also create categorical distinctions (i.e., in Sheng et al., a less than 5% difference between L1 and L2 language scores was considered “balanced”). Similarly, the self-report assessment tool developed by Anderson et al. (2018) can be used continuously or categorically; factor analysis was used to compute a bilingualism factor score, but Anderson et al. (2018) also provide recommended cut-offs for defining groups in their publicly accessible examiner materials. 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 important issue that emerges when reviewing this literature is variability in the contextual nature of the samples, including age, risk status, and languages spoken. The majority of the aforementioned studies have been with adults, with fewer focused on children (Archila-Suerte et al., 2016; Kim et al., 2018; Lonigan et al., 2018; Sheng et al., 2014; Vega & Fernandez) and even fewer with children identified as at-risk (Kim et al., 2018; Lonigan et al., 2018). Moreover, not all of the aforementioned studies utilized samples of Spanish-English speaking bilinguals (Anderson et al., 2018; Sheng et al., 2014; Yow & Li, 2015). In the present study, we consider these characterization issues among Spanish-English speaking middle school ELs with reading difficulties.

Finally, in reviewing these studies, a fifth important issue emerges, which is whether studies *distinguish* between language level/proficiency and balance. Many of the above 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 (a within-person distinction). However, it is possible for an individual to have equivalent knowledge of their L1 and L2, but to have low (or high) proficiency in both languages relative to expected norms for language skills (a between-person distinction). The difference between these two factors is not often made, likely because many of the samples being evaluated have L1 and L2 proficiency within the average range or higher. One exception is the study from Rosselli et al. (2016) which used a sample of undergraduates, where balanced bilinguals were classified into two groups, either “balanced-high” or “balanced-low;” however, they did not create two groups for their unbalanced group (i.e., “unbalanced English dominant,” “unbalanced Spanish dominant”) and thus were unable to fully categorize their participants on both proficiency and balance. The distinction between balance and language proficiency is especially important when considering the role of language in 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, or an unbalanced bilingual with lower proficiency in the language of instruction, as opposed to a balanced bilingual who is highly proficient in each language. Thus, in at-risk contexts, characterizing variability in bilingualism should encapsulate both proficiency and balance. For investigations of complex outcomes such as cognitive skills or academics, it would be important to clarify whether balance predicts performance above and beyond the impact of overall language proficiency, and this is a specific focus of the present study (in Chapter 2).

Despite the array of literature covered above, we are nonetheless only aware of a few studies that considered both language proficiency and balance in characterizing samples of Spanish-English speaking bilinguals (Lonigan, Goodrich, & Farver, 2018; Vaughn & Hernandez, 2018). For instance, in a sample of adults, Vaughn and Hernandez (2018) used a continuous metric of language proficiency which considered an additive combination of Spanish and English language composite scores as well as a function that provided a “boost” in score for individuals who were more balanced in their language abilities. Language ability was measured with two objective tests in each language: picture naming and passage comprehension. Their formula is shown in Equation 1.

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

Although this approach is informative in that it considers both balance and proficiency, there are many ways an individual could arrive at the same score; for instance, a higher score could either indicate that an individual was balanced with a moderate level of proficiency in each language or could show that an individual was unbalanced with a very high score in one of their languages. Thus, it is difficult to fully characterize an individual's language proficiency and balance based on this single score. Moreover, this study was conducted with adults.

In contrast, with a sample of preschoolers, a study from Lonigan et al. (2018) employed objective language measures in English and Spanish and used latent profile analysis to identify subgroups, which were then compared on early literacy skills. Two objective measures of language skills were used in each language, including tests of auditory comprehension (a complex receptive vocabulary measure) and expressive vocabulary. Nine distinct subgroups emerged and were characterized by patterns of L1 and L2 proficiency and balance. However, after evaluating the properties of the nine profiles and noting similarities among them, the researchers further partitioned participants into one of three "super" profiles: English Language Learners, Balanced Bilinguals, and Spanish Language Learners. Although these findings can help inform the characterization of language proficiency and balance in at-risk ELs, it is unclear how such groupings would emerge in a middle school sample of students who are further identified as struggling readers. Additionally, the selection of measures in the Lonigan et al. study was not informed by an evaluation of the dimensionality of language in their sample. It would also be of interest to investigate the extent to which a person-centered approach (i.e., latent profile analysis) maps onto language factors that result from a variable-centered approach (i.e., factor analysis). Furthermore, although the Lonigan et al. study did find that their subgroupings were related to measures of early literacy, they did not relate their subgroupings to any language-specific external measures. One such measure on which subgroupings could be compared is one of self-reported language usage, which is employed in this study.

Current Study

Taken together, the aforementioned studies demonstrate significant heterogeneity in characterizing bilingual samples in terms of language proficiency and balance, particularly with regard to method of assessment (i.e., objective vs. self-report), specific measures employed, approach used to define balance and/or proficiency, context of the sample, and whether the characterization included both L1 and L2 proficiency as well as consideration of balance. We are not aware of any studies that use such methods among middle school ELs who are also struggling readers, and are also not aware of studies that have systematically compared these various approaches. 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.

An evaluation of language in this context will begin with an investigation of the dimensionality of language through confirmatory factor analysis with a wide range of assessments (i.e., expressive, receptive, syntax, and semantics, in both English and Spanish).

Next, we will use latent profile analysis to determine which subgroups are present within our specific context 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., factor analysis) with the person-centered approach (i.e., latent profile analysis) in order to determine if these methods are convergent with one another. The identified latent profiles will then be compared on a closely related metric of proficiency and balance (i.e., the equation that appears in Vaughn and Hernandez, 2018) as well as a self-report measure of language usage. Finally, results from confirmatory models (i.e., factor loadings) can be used to inform a smaller set of measures which can then be evaluated to determine the extent to which they converge with the original factor scores, latent profiles, continuous metric, and self-report measure.

Hypotheses

1. Based on prior factor analytic work in bilingual samples of children (i.e., Gottardo & Mueller, 2009; Gray et al., 2018), we expect that our battery of objective language measures will demonstrate dimensionality with regard to language of assessment (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, it is predicted this will be along the dimension of semantics/syntax. Results from the best-fitting model will be used to create factor scores in English and Spanish that will be used as proficiency scores.
2. Given that the sample of students is at risk (ELs, struggling readers, from under-resourced schools), we expect latent profile analysis 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 anticipate that the latent profiles will differentiate according to English and Spanish proficiency factor scores computed through the best-fitting factor analytic model, demonstrating convergence between these approaches. This will be achieved by evaluating group differences across profiles on the English and Spanish factor scores using ANOVA. 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 expect that our latent profiles will differentiate according to scores 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. However, it is unclear how the other groupings may be characterized by this metric so we do not offer more specific hypotheses.
5. We predict that our latent profiles will also differ on a self-report measure of language usage; specifically, a higher level of balanced usage is expected among students in the balanced groups, a higher level of English usage is expected in the unbalanced-higher English proficiency group, and a higher level of Spanish usage is expected in the unbalanced-higher Spanish group.

6. In examining the use of a smaller set of measures for characterizing proficiency and balance, we do not offer specific hypotheses as there is not sufficient evidence from the literature on which to base predictions. However, we believe this is an important question to consider in order to inform which set of measures may allow for a more parsimonious approach to characterization of proficiency and balance.

Methods

Participants

Participants were 166 6^{th} and 7^{th} 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 represents a random subset of the larger Texas Center for Learning Disabilities (TCLD) sample (n=410) that received the language assessment battery for this project. Although all struggling readers in the TCLD were randomly assigned to either intensive reading intervention or business-as-usual (BAU) instruction, the current project is focused on pretest data collected prior to the onset of intervention, in order to mitigate any effects of intervention on the means of and/or the covariances among the language measures. In accordance with the TCLD project, inclusion criteria for all participants included: (1) enrolled in 6^{th} or 7^{th} grade; (2) identified as ELs or former ELs who have been re-designated as English proficient within the last five 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 TCLD project. Moreover, the majority of students in the middle schools served by the TCLD, as well as the local communities, reflect this demographic. Exclusionary criteria 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).

Although inclusionary criteria included being identified as a current or former EL, inspection of our final sample revealed that five students had never been identified as ELs. Because the purpose of this study was to characterize bilingualism in the specific context of ELs, we chose to remove these five students from our analyses; however, we note that our pattern of results was similar regardless of whether or not they were included. Thus, our final sample size was n=161.

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. 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 represented in the sample. Classrooms had, on average, six students represented, with six classrooms being represented by only one student.

Procedures

Recruitment involved obtaining permission from school districts to contact principals at a number of middle schools (in the context of the larger parent project). Teachers in grades 6 and 7 were then briefed about the study, provided information, and an opportunity to answer any questions. If interested and willing, informed consent letters were sent home to students' families. All examiners were trained by experienced assessment coordinators. All procedures were approved by the Institutional Review Boards of the University of Houston and the University of Texas at Austin.

All assessments were administered by trained, supervised data collectors. Data collectors were hired as part of the TCLD project and included bilingual data collectors to administer Spanish language assessments. Data collectors were trained over a three-week period and training involved formal review of examiner manuals as well as practice with test administration and scoring. Administrative staff members fluent in Spanish were available to assist in examiner training for portions of the assessment protocols conducted in Spanish. All data collectors were tested by project investigators before being approved to test in the schools.

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, & Shrank, 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, 2011) is a measure of *bilingual* receptive language and thus items are administered in Spanish and/or English. For standard administration of the test, items are first presented in one language (either Spanish or English, depending on which language the examiner believes to be dominant for that particular student). If correct, the student receives credit for that item. If incorrect, the same item is re-presented in the second language. However, for the purposes of this study we needed a score that reflected Spanish receptive vocabulary only; therefore, we administered each item in Spanish first. If the student answered incorrectly, the item was also administered in English. Thus, the standard score that results from standard test administration reflects overall receptive vocabulary in both English and Spanish. However, we computed a raw score that only included total number of items answered correctly when administered in Spanish. 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* 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, 2003) 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 from the English version and from .62 to .98 on subtests from the Spanish version. 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 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 have not been normed with the rest of the measure;

however, we computed an average score for each student such that a higher score indicates more English usage. The reliability within the present sample was .67; however, it could be argued that one item (Item 1) on this scale was not an index of language usage but rather an index of perceived relative proficiency. The item asks, "Which language do you know better?" with the same response choices as the other items (Mostly Spanish, Half Spanish/Half English, or Mostly English). When this item was removed from the scale, however, internal consistency did not improve (alpha = .59). However, in subsequent analyses we considered the full self-report measure, item 1 by itself, and the remaining 8 items by themselves in order to examine possible differences between results for the usage items and the perceived relative proficiency item.

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). There was non-normality noted on three measures, *Batería-III Memory for Sentences*, *Batería-III Picture Vocabulary,* and *ROWPVT-4*. Across these measures, eleven outliers were identified. Data reduction methods (factor analyses and latent profile analysis) utilizing the nine standardized language variables were conducted with and without these outliers and demonstrated the same pattern of results. Therefore, these outliers were retained in the final analyses.

As noted, age-based standard scores were not available for the Spanish measure of receptive semantics (*ROWPVT Bilingual Edition*) given how this measure was administered. Therefore, raw scores for all nine language measures were standardized in SAS with a mean of zero and standard deviation of one before being outputted to MPlus for factor analyses and latent profile analysis. Thus, we report both standardized raw scores and age-based standard scores in Table 1 but note that the standard score reported for the *ROWPVT Bilingual Edition* is the score obtained from typical test administration (which as noted correlated highly with our modified administration).

A variable-centered approach was used to test Hypothesis 1. Specifically, in order to evaluate whether the nine objective language tests can be explained by underlying latent factors, confirmatory factor 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). CFA models were tested in MPlus (Muthén & Muthén, 2012). Due to non-normality for three measures, maximum likelihood estimation with robust standard errors (MLR) was used across confirmatory models; however, we note that our pattern of results was the same regardless of whether MLR or maximum likelihood estimation was used.

Additionally, the type=complex option was used in MPlus across confirmatory models, as accounting for clustering of students within classrooms results in more accurate standard errors (Snijders & Bosker, 2011).

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), which reflects the standardized difference between the observed and predicted correlations, 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. The comparative fit index (CFI) was used as a comparative fit index. CFI values greater than .90 generally indicate good fit (Schumacker & Lomax, 2004).

A person-centered approach was used to test Hypothesis 2. Specifically, latent profile analysis (LPA) was used to evaluate whether students could be grouped according to their pattern of performance across the nine language tests. This approach differs from the variable-centered CFA models described above because LPA assumes that an underlying latent categorical variable divides a population into mutually exclusive latent profiles based on their pattern of performance on a range of continuous indicator variables. In contrast, CFA involves multiple continuously distributed latent factors based on correlations between indicator variables. Finally, 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 will also be computed to evaluate each model. The model entropy statistic ranges from 0-1 and provides an index of classification certainty; that is, how well each student's most likely profile membership corresponds to their actual profile membership as identified by the model. Higher entropy values indicate better classification certainty; prior studies have employed a cutoff of 0.80 (Hart et al., 2016; Lonigan, Goodrich, & Farver, 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. Higher entropy values correspond to higher average posterior probabilities. 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. Additionally, the final model should not contain any profiles containing less than 1% of the sample.

Resultant latent profiles 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 scores on a self-report measure of language usage (Hypothesis 5). Finding that latent profiles significantly differ across these metrics would provide construct validation for the LPA results, with pairwise comparisons illustrating specific differences among the profiles in expected directions. For example, a latent profile characterized by balance between Spanish and English, and average language proficiency in each language, was expected to correspond with proficiency factor scores falling in the average range for both L1 and L2. In contrast, an unbalanced latent profile characterized by average Spanish and low English was expected to correspond to average Spanish factor scores and low English factor scores. Similarly, since the Vaughn and Hernandez (2018) metric combines both proficiency and balance, a higher score on this metric was expected in a balanced latent profile with average Spanish and English language proficiency, whereas a lower score on this metric was expected for an unbalanced latent profile with average Spanish and English language proficiency. Since the self-report measure of language usage results in higher scores for English use, it was expected that latent profiles characterized by average English proficiency (i.e., a balanced profile with average English and Spanish, or an unbalanced profile with average English and low Spanish) would demonstrate higher scores on this self-report measure.

Finally, for Hypothesis 6, results from the best-fitting confirmatory factor model were inspected to determine whether a smaller set of measures may be appropriate for indexing L1 and L2 proficiency, which could then be further compared against findings from Hypotheses 3 through 5. This will be achieved by considering the factor loadings in the best-fitting factor model.

Results

Examination of descriptive statistics (Table 1) revealed that, on average, performance across both English and Spanish objective language measures fell significantly below normative age-based expectations. Specifically, age-based standard scores across all nine language tests fell at least one standard deviation below average. As a whole, the sample demonstrated an average standard score of 78.57 across English tests and an average standard score of 74.54 across Spanish tests. Since raw scores on language tests were standardized, we only report means and standard deviations for the age-based standard scores, though we do report correlations among both the raw standardized scores and among the age-based standard scores and note that these were nearly identical. As expected, bivariate correlations between tests within the same language were higher than correlations across languages. The five English tests demonstrated modest, significant relationships with one another, with correlations between age-based standard scores ranging from $r = .18$ to .39. The four Spanish tests demonstrated moderate to strong relationships with one another, with correlations between age-based standard scores ranging from $r = .36$ to .70. Cross-linguistic relationships were variable, with correlations ranging in both directionality and magnitude ($r = -.20$ to .31). Relationships between objective language tests and the self-report measure of language usage were generally higher for Spanish tests, with correlations between Spanish tests and the self-report measure ranging from $r = -.28$ to -.59. Relationships between English tests and the self-report measure ranged from $r = .01$ to .38.

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. Model 1 consisted of a single latent variable with all nine objective language measures but was a poor fit to the data, $\chi^2(26) = 135.05$, $p < .001$.

Model 2, a two-factor model differentiating between syntax and semantics measures, was examined next and also demonstrated poor fit, $\chi^2(25) = 133.43$, $p < .001$. Chi-square differences between Models 1 and 2 were examined using the Satorra-Bentler (2001) scaled (mean-adjusted) chi-square formula. There were no significant differences between these models, Satorra-Bentler χ^2 difference = 1.62, $p = .203$. Model 3, another two-factor model differentiating between 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. Model 4, a two-factor model differentiating between English and Spanish measures, provided strong fit, $\chi^2(25) = 55.09$, $p = .001$. Chi-square difference comparisons demonstrated that the English/Spanish model provided an improvement over the unitary 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, a further hypothesized model was tested that considered further distinctions between semantics and syntax within each language. Specifically, a four-factor model was tested including factors for English semantics, English syntax, Spanish semantics, and Spanish syntax. 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. This model provided 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 outputted from MPlus to SAS 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 3. 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 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, although we again note that the *ROWPVT Bilingual Edition* age-based standard scores were not employed in our analyses due to our modified administration. 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 categorized into Profile 1. This profile was characterized by balance between Spanish and English proficiency, with standardized raw scores on both Spanish and English being relatively higher than the other two profiles.

The average age-based standard score across all Spanish measures was 84.75, and the average age-based standard score across all English measures was 86.41.

Sixty-two percent (n=100) of the sample was categorized into Profile 2. 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 76.12 was noted across Spanish tests and an average of 75.15 was noted across English tests.

Twelve percent (n=20) of the sample was categorized into Profile 3. 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 45.75 across the four Spanish measures and an average English score of 79.56 across the five English measures.

The three latent profiles were exported from MPlus to SAS in order to evaluate their construct validity in subsequent hypotheses.

Convergence Between Variable-Centered and Person-Centered Approaches (Hypothesis 3)

In order to evaluate the convergence between the variable-centered (factor analysis) and person-centered (latent profile analysis) 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 latent profiles 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). 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)

We also utilized one-way ANOVA and Tukey-Kramer multiple comparisons tests to evaluate the extent to which the latent profiles differed from one another in expected directions 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 they were computed using the English and Spanish factor scores. We found support for this hypothesis, with latent profiles each differing significantly from one another in expected directions ($F = 106.35$, $p < .001$). Specifically, Profile 1 demonstrated the highest 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 performed significantly higher than Profile 3 ($p < .001$), with a large effect (d = 1.20).

Convergence Between Latent Profiles and Self-Report Measure (Hypothesis 5)

In order to evaluate the extent to which latent profiles differed on a self-report measure of language usage, we considered the self-report measure continuously with a one-way ANOVA test and found significant differences among the profiles, F = 12.27, $p < .001$. Tukey-Kramer multiple comparisons tests revealed some significant differences on the self-report measure in expected directions, though not all pairwise comparisons were significant as hypothesized. Specifically, Profile 3 reported significantly higher English usage relative to Spanish usage than Profile 1 ($p < .001$) and Profile 2 ($p < .001$), with large effects (d = 1.19 and d = 1.29, respectively). Profiles 1 and 2 did not differ significantly on the self-report measure (p = .841).

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 measures (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.

Since we noted that item 1 of the self-report measure indexed perceived relative language proficiency rather than language usage, we also considered this item separately to evaluate whether item response (Mostly Spanish, Half Spanish/Half English, Mostly Spanish) was associated with latent profile membership. The sample was sub-divided into three categories (unbalanced higher perceived Spanish proficiency, balanced proficiency, and unbalanced higher perceived English proficiency) and a chi-square test was used to evaluate convergence between the latent profiles and perceived relative proficiency. The chi-square test was significant, $\chi^2(4) = 25.68$, $p < .001$. Inspection of the ways in which students in the three latent profiles were distributed across the three perceived relative proficiency categories suggested some convergence between these two approaches, though not to the extent hypothesized. For instance, while we expected that most students in Profile 1 (characterized by balance between English and Spanish proficiency) would also report a balanced level of perceived relative proficiency, we found that only 13 out of 41 students in Profile 1 self-reported a balanced level of proficiency (16 reported higher Spanish proficiency, and 12 reported stronger English). The majority of the 100 students in Profile 2 (moderately unbalanced, higher Spanish proficiency) reported either a higher level of Spanish proficiency relative to English (n=42) or balanced proficiency (n=30), though many still reported a higher level of English proficiency (n=28). Consistent with hypotheses, the majority of the 20 students in Profile 3 (very unbalanced, English dominant) reported a higher level of English proficiency (n=17), a few reported balanced proficiency (n=3), and none reported a higher level of Spanish proficiency.

Finally, we also considered the remaining language usage items (full measure with item 1 removed) to determine if results were different when the perceived relative proficiency item was not included. We obtained the same pattern of results for this analysis as we did for the full self-report measure. Specifically, the one-way ANOVA was significant, $F = 9.37$, $p < .001$, indicating that the latent profiles differed on Items 2-9 of the self-report measure. Tukey-Kramer pairwise comparisons also demonstrated the same pattern of results as the full measure such that students in Profile 3 demonstrated a significantly higher level of English usage relative to both Profile 1 (p = .002) and Profile 2 (p < .001), with large effects (d = 0.99 and d = 1.09, respectively). Profiles 1 and 2 did not significantly differ from one another (p = .835).

Evaluating a Parsimonious Approach to Measurement of Language (Exploratory Aim 1)

In order to evaluate whether a smaller set of items could be used to characterize language proficiency and balance, we inspected factor loadings from the best-fitting factor model (see Figure 1). For both the English and Spanish factors, the measure of expressive semantics (*WJ-III Picture Vocabulary* and *Batería-III Picture Vocabulary*) provided the highest factor loadings (.59 and .87, respectively), while the *WJ-III Memory for Sentences* subtest provided an equally high factor loading on the English factor (.59), and provided the second highest factor loading on the Spanish factor (.78). Therefore, we chose to consider how each set of measures (the two expressive *Picture Vocabulary* scores and the two *Memory for Sentences* scores) mapped on to results obtained from the English and Spanish factor scores in order to evaluate more parsimonious approaches to characterizing proficiency and balance.

In order to evaluate whether the two *Picture Vocabulary* measures provided the same results as all nine measures, the three latent profiles (created using nine measures) were compared on English and Spanish *Picture Vocabulary* scores. A one-way ANOVA evaluating differences across the profiles on the English *Picture Vocabulary* scores was significant, F = 11.76, p < .001. Tukey-Kramer multiple comparison tests demonstrated the same pattern of scores as the English factor scores (Profile 1>Profile 3>Profile 2), although significant differences were only found for pairwise comparisons between Profile 1 and Profile 2 ($p < .001$) and Profile 2 and Profile 3 ($p = .005$), each with large effects ($d = 0.79$ and $d = 0.80$). Profiles 1 and 3 did not significantly differ on English *Picture Vocabulary* ($p = .995$). One-way ANOVA for differences across the profiles on the Spanish *Picture Vocabulary* scores was also significant, $F = 82.75$, $p < .001$, with significant pairwise comparisons found between all groups in the same direction as the results using Spanish factor scores (Profile 1>Profile 2>Profile 3). Specifically, Profile 1 demonstrated significantly higher Spanish *Picture Vocabulary* than both Profile 2 ($p = .003$) and Profile 3 ($p < .001$), with medium to large effects ($d = 0.65$ and $d = 2.82$, respectively). Profile 2 demonstrated significantly higher performance than Profile 3 ($p < .001$), with a large effect ($d = 2.38$). Next, the continuous metric integrating proficiency and balance was computed using only English and Spanish *Picture Vocabulary* scores. The correlation between this more parsimonious continuous metric and the original continuous metric was strong, $r = .78$ ($p < .001$). One-way ANOVA examined whether the three profiles differed in a similar manner on the continuous metric computed with *Picture Vocabulary* compared to the continuous metric computed with all nine measures. Profiles each differed significantly from one another, $F = 29.84$, $p < .001$, in the same manner as the continuous metric with all nine measures (Profile 1> Profile 2>Profile 3; all pairwise comparison p-values < .001, with large effects ranging from $d = 0.90$ to 1.91). Finally, we compared correlations between the factor scores and self-report metric to the *Picture Vocabulary* scores and self-report metric. English factor scores correlated $r = .24$ ($p = .002$) with the self-report measure, whereas English *Picture Vocabulary* correlated $r = .39$ ($p < .001$). Similarly, Spanish factor scores correlated $r = -.56$ ($p < .001$) with the self-report measure, whereas the Spanish *Picture Vocabulary* scores correlated $r = -.59$ ($p < .001$). Therefore, results utilizing *Picture Vocabulary* in each language provided the same pattern of results for indices of proficiency and balance as the full battery of language assessments.

The same set of analyses were conducted in order to evaluate the extent to which the two *Memory for Sentences* measures provided the same results at the full set of nine language measures. A one-way ANOVA evaluating differences across the profiles on the English *Memory for Sentences* scores was significant, $F = 36.25$, $p < .001$. Tukey-Kramer multiple comparison tests demonstrated the same pattern of scores as the English factor scores (Profile 1>Profile 3>Profile 2), although significant differences were only found for pairwise comparisons between Profile 1 and Profile 3 ($p < .001$) and Profile 1 and Profile 2 ($p < .001$), each with large effects ($d = 1.33$ and $d = 1.62$). Profiles 2 and 3 did not significantly differ on English *Memory for Sentences* ($p = .258$). One-way ANOVA for differences across the profiles on the Spanish *Memory for Sentences* scores was also significant, $F = 56.61$, $p < .001$, with significant pairwise comparisons found between all groups in the same direction as the results using Spanish factor scores (Profile 1>Profile 2>Profile 3). Specifically, Profile 1 demonstrated significantly higher Spanish *Memory for Sentences* than both Profile 2 ($p < .001$) and Profile 3 ($p < .001$), with large effects ($d = 0.94$ and $d = 2.11$, respectively). Profile 2 demonstrated significantly higher performance than Profile 3 ($p < .001$), with a large effect ($d = 1.51$). Next, the continuous metric integrating proficiency and balance was computed using only English and Spanish *Memory for Sentences* scores. The correlation between this more parsimonious continuous metric and the original continuous metric was strong, $r = .70$ ($p < .001$). One-way ANOVA examined whether the three profiles differed in a similar manner on the continuous metric computed with *Memory for Sentences* compared to the continuous metric computed with all nine measures. Profiles each differed significantly from one another, $F = 47.41$, $p < .001$, in the same manner as the continuous metric with all nine measures (Profile 1> Profile 2>Profile 3. All pairwise comparisons were significant with medium to large effects ranging from $d = 0.71$ to 2.11. Finally, we compared correlations between the factor scores and self-report metric to the *Memory for Sentences* scores and self-report metric. English factor scores correlated $r = .24$ ($p = .002$) with the self-report measure, whereas English *Memory for Sentences* correlated $r = .30$ ($p < .001$). Similarly, Spanish factor scores correlated $r = -.56$ ($p < .001$) with the self-report measure, whereas the Spanish *Memory for Sentences* scores correlated $r = -.44$ ($p < .001$). Therefore, results utilizing *Memory for Sentences* in each language also provided the same pattern of results for indices of proficiency and balance as the full battery of language assessments.

For a summary of results for each hypothesis, see Table 4.

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 who have been 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 are expected given an EL designation, what was striking was that Spanish scores were lower than English skills on average, and accompanied by wide variability. At the level of our hypotheses, results demonstrated that 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 and proficiency converged with objective measures, though not to the extent hypothesized. Similar results were obtained when using only measures of *Picture Vocabulary and Memory for Sentences*. These results help inform the selection of measures, as well as approaches to the characterization of bilingual students.

The Structure of Language Among Middle School ELs with Reading Difficulties: A Variable-Centered View

One of our primary aims was to evaluate the dimensionality of English and Spanish language processes in this population. Based on our review of the literature, we predicted that confirmatory factor models would support a two-factor structure consisting of one English factor and one Spanish factor. Our results clearly support this distinction 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 et al. (2006), who argued that a joint measurement model of English and Spanish tasks is needed in order 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 & Gutierrez-Clellen, 2009) but inconsistent with other bilingual studies that report high correlations between English and Spanish language factors (Castilla et al., 2009; Gray 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 Gray 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. In contrast, while the Simon-Cereijido and Gutierrez-Clellen (2009) study, whose results were consistent with those of the present study, did not report age-based standard scores, they noted that over one third of their sample had been diagnosed with a language delay.

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 $r = .70$. 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. Problems related to construct validity would suggest that a different set of norms, or perhaps a different set of measures, may be warranted for evaluation of English language skills in this population. 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, however, 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 students’ 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. However, it is also possible that this was an age-related effect and that we would have seen further differentiation between language skills had we evaluated our sample at an earlier timepoint. In support of this idea, Gray et al. (2018) found that Spanish syntax differentiated from Spanish vocabulary in their sample of kindergarteners, and the factors correlated at $r = .56$. Another 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. 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. Moreover, effective application of grammatical rules and syntactic structure presupposes some level of understanding of the semantic content of the material. 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. Future work may wish to employ syntax measures that are weaker in their semantic demands such that individual differences are due to true differences in syntactic abilities (e.g., the Token Test; De Renzi & Vignolo, 1962). Another approach could be to use an experimental paradigm such as an artificial grammar learning task (Chomsky & Miller, 1958), though this task has been criticized due to lack of generalizability to real-world settings.

It is also possible that we did not find differentiation of semantic and syntactic skills, particularly in Spanish, because of low language proficiency. Specifically, as intercorrelations between tests increase, and as overall abilities go down, profiles of abilities tend to flatten (Fletcher, Denton, & Francis, 2005; Morris, Fletcher, & Francis, 1993). While we did not have subgroups of students in our sample with high average or even average language proficiency that emerged from our latent profile analysis, it is possible that such students may exhibit greater differentiation between semantics and syntax skills. Perhaps such differentiation may even have emerged in our highest performing subgroup, Profile 1, but a factor analysis would not be practicable with such a small sample (*n* = 41). Importantly, studies in monolinguals that support a distinction between semantics and syntax at the middle school level have used samples with well-developed language skills (Foorman et al., 2015; Lonigan & Milburn, 2017; Tomblin & Zhang, 2006). While we were unable to differentiate whether the pattern of results is more related to the measures themselves versus the sample characteristics (or their interaction), we emphasize the significance and novelty of our current findings, which improve our understanding of language processes in this at-risk population. Importantly, the lack of differentiation between semantics and syntax as suggested by our factor models does not mean that these language processes are not each important in their own right.

Classifying Students by Proficiency and Balance: A Person-Centered View

In addition to using a variable-centered approach to inform English and Spanish factor scores as continuous indices of language proficiency, another primary aim of this study was to evaluate a person-centered approach to classification. As expected, latent profile analysis demonstrated that the students in our sample could be sub-grouped based on language proficiency levels as well as balance between English and Spanish skills, although we found that our students were subdivided into three profiles rather than our hypothesized four profiles. 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, it is important to also consider the age-based standard score performance across the profiles. These two different sets of scores 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 approaches were consistent for Profiles 1 and 3 (for instance, Profile 1 was balanced and demonstrated the highest English and Spanish scores regardless of whether we looked at standardized or age-based standard scores), 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 more 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 we again note that there may be practical reasons for using and interpreting age-based standard scores on these tests in our sample.

Findings from the latent profile analysis also highlight the wide variability in Spanish skills relative to English in this sample, with students in Profile 3 performing in the impaired range, on average, across all five Spanish measures. Although the three profiles also differed in their English language skills, the wider variability in Spanish skills may be due to the fact that these students share an English-speaking classroom environment, making their English use/exposure somewhat more homogeneous, whereas they may differ in the extent to which they use/are exposed to Spanish in their home and community environments. Although this data was unavailable to us, more information regarding the students' instructional history with regard to 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. Specifically, we found that the latent profiles differed in expected directions on the English and Spanish factor scores as well as the continuous metric integrating balance and proficiency which was computed using the factor scores (Vaughn & Hernandez, 2018). We note, however, that one limitation of interpreting these results is that the latent profiles were used in a deterministic rather than a probabilistic manner, as profile membership was exported from Mplus categorically rather than probabilistically. This limitation could be addressed in future studies through the use of weighting or resampling techniques.

Convergence of Objective Language Measures with a Self-Report Measure

We chose to use a self-report measure as a construct validity target because a primary aim of this study was to better understand how information gathered from self-report tools converges with or is complementary to objective measures. In partial support of our hypothesis, we found that Profile 3 reported significantly lower Spanish usage relative to English usage than Profiles 1 and 2; however, Profiles 1 and 2 did not differ significantly from one another. Thus, while the self-report measure did track with the overall trend of results for the objective measures, it was not able to differentiate between students in profiles that were characterized by differences in English proficiency, Spanish proficiency, and balance.

While the correlation we found between the Spanish factor score and the self-report measure was moderate, $r = -.56$, we were surprised by the low correlation between the English factor score and the self-report measure, $r = .24$. This finding again may point to possible issues of construct validity of the English measures but must also be considered in light of the weak reliability of the self-report measure. It is also possible that the scaling of this measure was not appropriate as response options ranged from “All Spanish” to “All English” rather than having separate sets of items for degree of English usage and degree of Spanish usage. However, it is important to note that to our knowledge, there is currently no self-report tool designed to measure language usage or proficiency in bilingual children. Moreover, the items we had available for indexing self-reported language usage and proficiency were not explicitly designed for this purpose. The items, which are part of the *ROWPVT-Bilingual Edition*, are designed to be administered at the beginning of the test in order to inform the examiner’s decision about whether to begin administration in Spanish or in English (examiners are told to begin test administration in what they perceive to be the student’s dominant language).

We also noted that one item on the self-report measure appeared to assess perceived relative proficiency rather than language usage. However, after considering this single item, the remaining usage items, and the full measure through three separate analyses, the same pattern of results emerged such that Profile 3 differed from the other two profiles, but Profiles 1 and 2 did not differ significantly from one another. While these findings suggest that the relationship between self-report and objective measures is comparable regardless of whether usage or perceived relative proficiency is measured, it is important that future work evaluate this more comprehensively, and with more reliable instruments. Additionally, items that ask about perceived proficiency level in *each* language (rather than relative proficiency in one language compared to the other) may provide additional information which may or may not converge with these other types of items.

Parsimonious Evaluation of Language

A final aim of this study was to evaluate whether using a smaller set of measures would produce the same pattern of results as a larger battery of objective language tests. Such investigation holds clear practical implications, since the use of single tests in each language is a common approach used for characterizing bilinguals in this literature (i.e., Gollan et al., 2012; Sheng et al., 2014; Tomoschuk et al., 2019). We found that the *Picture Vocabulary* subtest in English and Spanish demonstrated the same pattern of results as all nine objective measures, and also that the *Memory for Sentences* subtests demonstrated the same general pattern of results. These findings could be used to support the practice of using single tests in L1 and L2 in future work seeking to characterize bilingual samples. As noted above, it is likely that our syntax measures, including *Memory for Sentences*, were capturing both syntactic as well as semantic knowledge. Thus, this measure may be preferred when only employing a single measure, as it may provide a more comprehensive index of language skills than *Picture Vocabulary*. Importantly, however, it often may not be appropriate to rely on only two indices of language skill. For instance, as noted above, further work is needed to clarify the extent to which language subskills may be separable in confirmatory factor models as overall proficiency levels increase, and there is clinical and educational relevance for considering different aspects of language skill and how they may differentially relate to achievement and functional outcomes.

Resolving Core Issues Regarding Characterization of Proficiency and Balance in Bilingual Samples

In considering questions about the dimensionality of language among bilinguals, possible subgroupings of students, relations between objective and self-report measures, and consideration of a more parsimonious approach, the overarching purpose of the current study was to take steps towards resolving core issues we noted in our review of the literature regarding the characterization of bilingual samples. With regard to the issue of variability across studies with regard to reliance on self-report versus objective measures of language, the results of the current study suggest that reliance on self-report alone is unlikely to provide a full picture of a students' English and Spanish skills, particularly if there is limited evidence regarding the psychometric properties of the instrument for the specific population of interest. Similarly, a related and somewhat overlapping issue is variability with regard to approaches used to define balance in the literature. We used objective tests to evaluate both variable-centered and person-centered approaches to characterize balance and found that these converged with one another in expected directions, but the self-report measure only weakly mapped onto these findings, again highlighting the limits of these measures for the population studied here. Importantly, by considering metrics that integrate both balance and proficiency, this work opens the door for future studies to evaluate phenomena that are commonly examined in bilingual research, such as the possible role of balanced bilingualism in conferring a bilingual advantage, in at-risk populations. It is particularly important to consider both proficiency levels and balance in such at-risk samples because the impact of being a balanced bilingual may well depend on the language proficiency levels of the individual.

In addition to these measurement-related issues, we noted that the vast majority of studies that characterize bilinguals in terms of proficiency and/or balance have been with adults whose language proficiency levels are within the average range or higher, and thus it is difficult to generalize findings from this literature to at-risk populations of children. This study has taken an important first step in characterizing the heterogeneity of language abilities among middle school ELs with reading difficulties, and such information should be used as a jumping-off point for future work which examines how variability in the bilingual experience impacts important outcomes in this population.

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 impairment 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. Moreover, as noted, there is some utility to 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 weak reliability found for the self-report measure, which may also be related to problems with the scaling of this measure as noted above, 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 (i.e., Anderson et al., 2018). In doing so, it is also important for future work to consider convergence between reports from multiple informants of the students' language usage and proficiency (i.e., teachers, parents). It is also possible that our chosen metric for self-report (i.e., Likert scale with three responses choices, Mostly Spanish, Both, or Mostly English) was not sufficient for capturing the variability with which students utilize their languages. Recently, Gullifer and Titon (2019) have proposed the concept of language entropy, which takes into account individual differences in the social and interactional context of language usage. For instance, some bilinguals may be compartmentalized in their language usage (i.e., use one language for some contexts, and one language for others), while others may integrate their language usage within one context. Gullifer and Titon (2019) developed a metric for computing language entropy which can be applied to self-report data such as that gathered in the current study.

A final limitation may be the somewhat restricted range of language proficiency in our sample, though the purpose of our study was to investigate characterization of language in this at-risk sample, 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 is represented in the current study. Such investigation may clarify some of the questions raised in the current work; for instance, results may demonstrate increasing multidimensionality along the syntax/semantics dimension for students with higher levels of proficiency.

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 in order 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 in a bilingual sample of middle school students. Our findings reflect the multidimensionality of language in this important and understudied sample (along the English/Spanish dimension), a three-profile group structure, convergence between variable-centered and person-centered methods, partial support for the use of self-report tools, and support for a parsimonious approach to measurement. 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.

Chapter 2: The Roles of L1 and L2 Language Proficiency and Balance in Reading Outcomes among Middle School English Learners with Reading Difficulties

Benefits associated with bilingualism have been a focus of study as the proportion of the population that speaks more than one language has continued to increase (Ortman & Shin, 2011). Specifically, investigations of the bilingual advantage hypothesis are common within the domains of psychology, linguistics, and neuroscience. This hypothesis argues that navigating the use of two or more languages requires greater cognitive resources than speaking only one language, therefore resulting in higher levels of executive control, switching, cognitive flexibility, and inhibition among bilinguals compared to monolinguals (i.e., Carlson & Meltzoff, 2008; Bialystok and Martin, 2004; Bialystok et al., 2008; Morales et al., 2013). However, many groups have found mixed or non-significant results for this hypothesis (i.e., Anton et al., 2014; Gathercole et al., 2014; Paap & Greenberg, 2013; Paap & Sawi, 2014). However, these inconsistent findings may be related to the fact that bilinguals are typically considered in aggregate rather than considering important individual differences among bilinguals.

To take the bilingual advantage hypothesis a step further, it has been argued that the *balance* of an individuals' bilingualism (i.e., the extent to which individuals have similar L1 and L2 proficiency and/or usage) may independently explain differences in cognitive outcomes among bilinguals (Verreyt, Woumans, Vandelanotte, Szmalec, & Duyck, 2016; Yow & Li, 2015). The basis for this argument is that the cognitive benefits associated with bilingualism derive from frequent switching and inhibition between languages. Therefore, the greater the balance, the more frequent would be the regular practice of navigating between two (or more) languages, and thus bilingual benefits may be amplified. There is some emerging support for the role of balance in executive function (EF) outcomes (Bialystok & Barac, 2012; Vega & Fernandez, 2010; Yow & Li, 2015). However, as reviewed in Chapter 1, studies that consider balance vary in the extent to which they consider balance in language *proficiency*, balance in language *usage*, and overall L1 and L2 *level(s)*. Additionally, much of our knowledge about the ways in which bilingualism impacts important outcomes comes from studies with adult samples whose L1 and L2 abilities are generally well-developed; in contrast, studies that use at-risk samples of children are less common. However, some emerging evidence suggests that bilingualism may also confer advantages in at-risk developmental contexts (White & Greenfield, 2016).

Furthermore, while studies testing the bilingual advantage have focused on differences in EF, it is important to extend these findings by considering other outcomes that are likely to be impacted by variability in bilingualism. Reading is a language-based skill (Vellutino, Fletcher, Snowling, & Scanlon, 2004) that is also rooted in domain general cognitive abilities including EF (Butterfuss and Kendeou, 2017; Cirino et al., 2019), and has a strong evidence base on which to inform intervention (Scammacca et al., 2016). Thus, it is important to better understand how variability of the bilingual experience could potentially be leveraged to improve reading outcomes, particularly in at-risk populations. Therefore, the purpose of the current study is to examine the extent to which these three factors (degree of balanced proficiency, L1 and L2 language level, and self-reported language usage) predict reading outcomes among an at-risk sample of children.

The Bilingual Advantage and Evidence for the Role of Balance

A number of studies have found support for a bilingual advantage across the lifespan (Bialystok, 1999; Bialystok & Martin, 2004; Bialystok et al., 2008; Carlson & Meltzoff, 2008; Morales et al., 2013). For example, Bialystok and colleagues (2008) found that bilinguals had stronger performance on an interference inhibition task (Stroop) compared to monolinguals. Similarly, in a sample of native English-Spanish kindergarteners compared to monolinguals and English speakers enrolled in a Spanish immersion program, Carlson and Meltzoff (2008) found support for a bilingual advantage on a task requiring management of conflicting attentional demands. A meta-analysis of available evidence for the bilingual advantage in adults and children conducted by Adesope et al. (2010) concluded that existing evidence supports an advantage for bilinguals across multiple measures of skills that could be considered to fall under the domain of EF, with effect sizes in the small to moderate range. For instance, Adesope et al. (2010) reported an effect size of $g = 0.33$ for measures of metalinguistic and metacognitive awareness, and an effect size of $g = 0.52$ for measures of abstract and symbolic representation, attentional control, and problem solving.

In contrast, a number of studies have found mixed or no evidence for this theory. For instance, EF advantages may emerge only on certain tasks or in certain conditions. Costa et al. (2009) found that bilinguals performed better on a flanker task when high monitoring resources were required to ignore distracting information whereas no advantage was found when only low monitoring resources were needed. A number of recent studies have found no evidence of a bilingual advantage (Anton et al., 2014; Gathercole et al., 2014; Paap & Greenberg, 2013; Paap & Sawi, 2014). For example, in a sample of undergraduates, Paap and Greenberg (2013) compared primarily English-Spanish speaking bilinguals to monolinguals on fifteen indicators of EF and found no significant group by treatment interactions in favor of a bilingual advantage. Similarly, in a sample of third through sixth grade children, Anton et al. (2014) found no differences on a test of attentional control between Spanish-Basque speaking bilinguals and Spanish-speaking monolinguals.

In order to reconcile these mixed findings, it has been suggested that balance between L1 and L2 is important to consider in investigations of the bilingual advantage (Bialystok & Barac, 2012; Vega & Fernandez, 2011; Verreyt et al., 2016; Yow & Li, 2015). Because EF benefits derive from the regular practice of using and navigating between two languages (which requires monitoring, attentional, and inhibitory resources), then it is reasonable to propose that a bilingual individual who is more balanced in their language usage and/or language proficiency may have conferred advantages on cognitive tasks involving EF compared to a bilingual who is less balanced. In support of this idea, Yow and Li (2015) posited that "if frequent practice of controlling and attending to the appropriate language system confers general advantage in executive control tasks, then balanced bilingualism…would critically affect the development of executive control skills in bilinguals" (p. 3). However, empirical investigations regarding the impact of balance on important outcomes are limited, particularly among at-risk samples of children, and particularly in areas outside of EF.

Yow and Li (2015) recently found support for the role of language balance among an adult sample of English-Mandarin speakers. They found that balanced *usage* predicted performance on EF tasks of inhibition and set-shifting, and balanced *proficiency* predicted performance on a set-shifting task. Similarly, Verreyt et al. (2016) also found support for considering the role of balance in cognitive outcomes. Specifically, in a sample of adult Dutch-French speaking adults, they found stronger performance on two executive control tasks for balanced bilinguals who reported frequently switching between their two languages when compared to unbalanced bilinguals and balanced bilinguals who reported low levels of switching between languages.

Only a few studies have considered the role of balance in outcomes for children. For instance, Vega and Fernandez (2010) compared EF tasks of set-shifting and inhibition among Spanish-English speaking third and fourth grade children classified as “more balanced” and “less balanced.” They found that their “more balanced” group performed significantly better on the set-shifting task than the “less balanced” group, but no differences were found on the inhibition task. Additionally, although not with Spanish-English speaking bilinguals, Bialystok and Barac (2012) found some support for balance in L1 and L2 in a sample of second and third grade English and Hebrew-speaking children. Even when also considering overall language proficiency, they found that degree of balance predicted performance on a flanker (inhibition) task but not on a task-switching test. Language proficiency in both languages was in the average range for this sample, and therefore it is unclear if results would hold in an at-risk sample.

Notwithstanding the variability across studies in the way that balance is defined and how it is integrated with proficiency level (see Chapter 1 for a discussion, pp. 11 to 17), and despite the emerging evidence above, we are not aware of *any* studies that consider how balance is related to outcomes in middle school aged Spanish-English speaking students or among bilinguals with reading difficulties.

Extending the Bilingual Advantage Hypothesis to Reading in an At-Risk Context

The aforementioned review focuses on the emerging evidence for the role of balance in typical samples of adults, where the outcome is EF. However, it is also important to investigate the possible affordances of bilingualism among at-risk samples of children because doing so may help uncover potential intervention targets. One such population is middle school English Learners (ELs) with reading difficulties. In this context, reading is a novel and important outcome to consider when evaluating the impact of bilingualism. The review below seeks to inform hypotheses relating proficiency and balance to various reading outcomes by integrating findings from the well-established developmental reading literature (which focuses on the role of language) with those of the bilingual literature where reading has been investigated.

English Learners and Reading Achievement

English Learners (ELs) are an at-risk group of students in the US, and thus it is important to better understand the ways in which the bilingual experience impacts important outcomes such as reading. In the US, about 10% of students are categorized as ELs (Kena et al., 2015), a broad term which refers to students with low English proficiency (Genesee, Lindholm-Leary, Saunders, & Christian, 2005) who also speak a non-English language. Thus, ELs can be distinguished between other subgroups of bilinguals in that they are identified based on difficulties in English. The majority of ELs in the US speak Spanish as a first language (Halle, Hair, Wandner, McNamara, & Chien, 2012; Hindman & Wasik, 2015; Passel, Cohn, & Lopez, 2011). ELs are at higher risk for academic difficulties than monolingual students (Francis, Rivera, Lesaux, Kieffer, & Rivera, 2006; Hemphill & Vanneman, 2011), particularly if they have additional sociodemographic risk factors such as low socioeconomic status (Francis et al., 2006). Moreover, there is a dearth of research that focuses on learning difficulties and underlying cognitive predictors among ELs at the middle school level. Considering the extent to which L1 and L2 proficiency as well as balance relate to English reading outcomes in this population is an important next step in filling these research gaps regarding bilingualism, as greater understanding might inform future intervention for middle school ELs who are also struggling readers.

The Role of Language in Reading

Extending research on the bilingual experience to examine how variability in balance and proficiency relates to reading must also consider the vast prior work in broader developmental contexts regarding the role of language in reading achievement. For instance, the Simple View of Reading posits that reading achievement is influenced by a combination of decoding skill (single word reading) and oral language comprehension (Gough & Tunmer, 1986). Studies with ELs have also found support for this theory (Farnia & Geva, 2013; Geva & Massey-Garrison, 2013; Proctor, Carlo, August, & Snow, 2006), though these studies were conducted with late elementary school students rather than middle school students. Within the framework of the Simple View, oral language skills appear to play a stronger role in comprehension than decoding ability as students get older (Catts, Hogan, Adlof, & Barth, 2003). Thus, particularly among older samples of students (i.e., middle school), investigations focused on the role of oral language in reading have the potential to inform effective intervention approaches. With regard to theoretical frameworks for reading, it is also important to note that other models go beyond language by recognizing a wider range of cognitive processes (Perfetti & Stafura, 2014; van den Broek, Young, Tzeng, & Linderholm, 1999), and executive function in particular is one such domain general cognitive skill which has been implicated (Butterfuss & Kendeou, 2017; Cirino et al., 2019; Follmer, 2017).

Oral language skills broadly refer to one's understanding and use of oral language, which includes vocabulary and syntax; however, the extent to which language is best represented as a unidimensional or multidimensional construct throughout development is unclear, particularly among bilinguals (see Chapter 1 for discussion of this issue, pp. 8-11). In order to evaluate the dimensionality of language in the context of middle school ELs who are struggling readers, the first part of this study evaluated the factor structure of language using a wide range of objective language measures in English and Spanish. We found that language was best represented as two factors, one English and one Spanish.

The extent to which L1 language skills predict L2 reading above and beyond the contribution of L2 language skills is understudied, particularly at the middle school level and in among at-risk students. However, some work has considered these relationships among English-Spanish bilinguals more generally (Miller et al., 2006; Proctor, Carlo, August, & Snow, 2006; Proctor, Harring, & Silverman, 2017; Swanson, Orosco, & Lussier, 2011). For instance, Miller et al. (2006) found that a Spanish oral language composite consisting of semantic, syntactic, fluency, and discourse measures predicted English passage comprehension and word reading efficiency scores above and beyond the contribution of an English language composite in a large cross-sectional sample of kindergarten to 3rd grade Spanish-English bilingual children. Moreover, there was a stronger role of Spanish language for English passage comprehension relative to English word reading efficiency. In a sample of Spanish-English bilingual students in fourth grade, Proctor et al. (2006) evaluated the role of Spanish language skills in English reading comprehension while controlling for language of instruction, English decoding, and English oral language proficiency and found a significant main effect of Spanish vocabulary knowledge. Another study from Proctor et al. (2017) used a cohort-sequential longitudinal design to estimate growth trajectories from second through fifth grade and found that Spanish syntax predicted fifth grade English oral language skills and English reading comprehension.

In contrast, some studies have found that English language skills fully account for variability in English reading outcomes among Spanish-English bilinguals; for instance, Nakamoto et al. (2008) used structural equation modeling to relate Spanish and English oral language factors to English reading comprehension measured at the end of sixth grade. They found that the contributions of English decoding and oral language supplanted those of Spanish decoding and oral language in the prediction of English reading skills. These studies, however, did not select their samples for reading difficulties and did not consistently compare the contributions of L1 and L2 language skills to various reading outcomes (i.e., word reading accuracy, fluency, and comprehension). Moreover, this work has not also considered the role of balance between L1 and L2 in reading outcomes, and has not considered how a self-report measure of language may provide complementary information regarding the role of language in reading.

Reading as a Target for Bilingual Advantage Research

That reading is strongly rooted in language makes it a highly relevant outcome to consider in the bilingual context. Moreover, reading is also influenced by domain general cognitive abilities, including EF (Butterfuss & Kendeou, 2017; Cirino et al., 2019). It is highly relevant that aspects of EF often implicated in the bilingual advantage, including inhibition and shifting, are also thought to contribute to reading skill. For instance, in their conceptual evaluation of the roles of various EF processes in reading comprehension, Butterfuss and Kendeou (2017) concluded that inhibition is needed to suppress irrelevant information during reading comprehension tasks and shifting may be necessary for integrating lower-level material such as semantic and phonological information and applying it to comprehension of text. In addition to reading comprehension, EF has also been found to predict other reading outcomes including word reading accuracy and word reading fluency (Cirino et al., 2019). Moreover, it is likely more fruitful to target reading than EF, particularly where intervention is concerned, given the strong evidence base that informs effective reading interventions (Scammacca et al., 2016). In contrast, evidence for EF intervention is underdeveloped and intervening on such domain general abilities does not transfer to important functional outcomes such as academic achievement (for meta-analytic reviews on this topic see Kassai, Futo, Demetrovics, & Takacs, 2019; Melby-Lervåg, Redick, & Hulme, 2016).

A few studies have evaluated a possible bilingual advantage for literacy (Bialystok, Luk, & Kwan, 2009; Bialystok, Majumder, & Martin, 2003; Grimm, Solari, Gerber, Nylund-Gibson, & Swanson, 2019; Lonigan, Goodrich, & Farver, 2018). For example, Bialystok et al. (2009) compared four groups of 1st graders (one monolingual group and three bilingual groups of different language backgrounds including Spanish-English speakers) on decoding and phonological awareness tasks. They found stronger outcomes for all three bilingual groups compared to the monolingual group, especially for bilinguals who spoke two alphabetic languages. However, these studies were not conducted with struggling readers and they focus on pre-literacy skills rather than reading outcomes such as word reading accuracy, fluency, and comprehension. Additionally, these studies did not consider the role of balance between L1 and L2 in literacy-related outcomes.

In a sample of Spanish-English speaking preschoolers, Lonigan et al. (2018) used four objective measures of language and latent profile analysis to create subgroups of students characterized by differences in L1 and L2 proficiency as well as balance. Subgroups were then compared on English preliteracy outcomes. Although their latent profiles did demonstrate different patterns of performance across preliteracy assessments, their findings showed limited support for a bilingual advantage related to balance, as better English preliteracy skills were associated with stronger English language abilities. However, this study was with younger students and did not consider a range of reading outcomes.

In a sample of second grade Spanish-English speaking bilinguals from predominantly low socioeconomic households, Grimm et al. (2019) conducted a latent profile analysis with measures of word reading, expressive language, and receptive language in both languages. Three profiles emerged: a balanced profile and two unbalanced profiles, both of which had English skills in the average range but were differentiated by Spanish proficiency (either average or below average). In relating these profiles to reading comprehension, they found that students with the balanced profile outperformed students with unbalanced profiles; importantly, however, students in the balanced profile demonstrated somewhat higher scores across all measures relative to the other profiles. Results from this study suggest that having a balance between L1 and L2 may confer some advantage for reading, and that this advantage is present even in a low SES bilingual sample, though Grimm et al.'s investigation did not distinguish between balance and proficiency. However, this study was also with younger students, did not select language measures based on an investigation of the dimensionality of language, and did not also employ self-report measures of language. The extent to which these findings would hold in a middle school sample with identified difficulties in reading is an interesting and novel question that holds potential for informing theory and practice related to the bilingual advantage as well as reading achievement in this important at-risk context.

Current Study

The aforementioned corpus of studies suggests that investigations of the impact of bilingualism on important outcomes may consider three important factors: proficiency level, balance between language proficiency, language usage. In an at-risk context such as middle school ELs with reading difficulties, an important next step in advancing bilingual research is to evaluate the ways in which these factors relate to reading achievement, in either differential or complementary fashion. Therefore, the purpose of the current study is to evaluate how these three factors relate to L2 reading outcomes in a sample of middle school ELs with reading difficulties. The specific approaches to indexing proficiency, balance, and self-reported language usage were determined based on findings from Chapter 1, which are summarized below.

Summary of Chapter 1 Results

Chapter 1 utilized a range of objective language measures and evaluated both variable-centered (i.e., factor analysis) and person-centered (i.e., latent profile analysis) approaches to characterizing the sample in terms of L1 and L2 proficiency and degree of balance. These methods were then compared with one another and to additional metrics of language proficiency and balance, including a self-report measure. These findings indicated that Spanish and English proficiency can be represented by Spanish and English factor scores derived from the best-fitting confirmatory factor model using all nine objective language measures. Importantly, we evaluated the language levels of our sample both relative to one another and relative to age-based normative standards. Our sample fell within the low average range, and at least one standard deviation below normative expectations for their age level, across all nine language measures administered. However, significant variability was noted across the sample.

With regard to metrics that integrate both proficiency and balance, findings from Chapter 1 indicate that students in this sample can be classified into three latent profiles which are characterized by differences in L1 and L2 proficiency levels as well as balance. Specifically, students in Profile 1 were characterized by balance between L1 and L2 and relatively high levels of language proficiency in each language compared to students in the other profiles. Students in Profile 2 were characterized by a moderate level of imbalance between their L1 and L2 skills, with Spanish levels higher than English levels. Relative to the other profiles, Profile 2 had the lowest English proficiency levels and their Spanish levels were lower than Profile 1 but higher than Profile 3. Finally, Profile 3 was characterized by imbalance between L1 and L2, with Spanish levels falling significantly lower than English levels. Relative to the other profiles, Profile 3 had the lowest Spanish proficiency levels and their English levels were lower than Profile 1 but higher than Profile 3.

Chapter 1 also found convergence between the variable-centered confirmatory factor analytic approach and the person-centered latent profile analytic approach. Specifically, latent profiles were characterized by differences on the Spanish and English factor scores in expected directions. Additionally, findings from Chapter 1 indicated that a single continuous metric can also be used to index both proficiency and balance, as scores on this metric converged with the latent profiles in expected directions. With regard to the self-report measure of language usage, findings from Chapter 1 revealed a modest correlation between usage and English proficiency, and a moderate correlation between reported usage and Spanish proficiency. The latent profiles also differed on the self-report measure in expected directions, though not to the extent hypothesized (not all pairwise comparisons were significant); moreover, psychometric properties of the self-report metric were weak, and we noted that one item appeared to measure a different construct (perceived relative language proficiency).

Based on results from Chapter 1, the current study utilizes Spanish and English factor scores as indices of language proficiency, and both the latent profiles as well as the continuous metric as variables that integrate proficiency and balance. We also consider the possible role of self-reported language usage in reading outcomes, as well as the role of perceived relative language proficiency, though note that future studies should consider these relationships with a more reliable instrument.

Hypotheses

1. We predict that English language *proficiency* will have a larger impact on English reading achievement compared to Spanish language proficiency, when proficiency is indexed by English and Spanish factor scores obtained from Chapter 1. We expect this pattern for all types of reading examined (word reading accuracy, fluency, and comprehension).
2. In addition to indices of language proficiency, we expect that objective metrics integrating both proficiency and *balance* will be related to reading outcomes. We are not aware of any prior studies that evaluated the extent to which such metrics differentially predict various reading skills; however, we predict that latent profiles and the continuous metric of balance and proficiency may account for a higher proportion of variability in reading comprehension relative to word reading and fluency, as this task requires a higher level of cognitive resources.
3. We expect that scores on a self-report measure of *language usage* will also demonstrate an incremental contribution to reading outcomes (over the factors of

Hypotheses 1 and 2). Because the nature of this relationship is understudied with regard to reading, we do not present specific hypotheses; however, we believe it is possible that either a higher level of L2 (English) usage may be related to reading, or that a greater balance between L1 and L2 usage may be related to reading. Further, it is possible that the nature of these results may differ for each reading outcome; for instance, a higher level of L2 usage may be more important for word reading accuracy, but balanced usage may confer an advantage for reading comprehension. Moreover, we also predict that an index of perceived relative language proficiency will relate to reading, but the nature of this relationship is also unclear.

Methods

Participants

This sample represents a random subset of the larger Texas Center for Learning Disabilities (TCLD) sample (n=410) that received the language assessment battery for this project. Although all struggling readers in the TCLD were randomly assigned to either intensive reading intervention or business-as-usual (BAU) instruction, the current project is focused on pretest data rather than the intervention component. In accordance with the TCLD project, inclusion criteria for all participants includes: (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 five years; (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 is necessary to reduce heterogeneity of the sample for the epigenetics portion of the larger TCLD project. Moreover, the majority of students in the middle schools served by the TCLD, as well as the local communities, reflect this demographic. Exclusionary criteria included: (1) a sensory disorder that precludes participation in the assessment and intervention protocols; and (2) participation in an alternative curriculum (i.e., life skills course).

Although inclusionary criteria included being identified as a current or former EL, inspection of our final sample revealed that five students had never been identified as ELs. Because the purpose of this study is to characterize bilingualism in the specific context of ELs, we chose to remove these five students from our analyses; however, we note that our pattern of results was similar regardless of whether or not they were included. Thus, our final sample size was n=161. However, we note that the reading comprehension measure was missing for one student, so those regression models were conducted with a sample of n=160.

Forty-eight percent of students were in 6^{th} grade and 41% were female. The mean age of the students was 12.5 years (*SD* = .75 years). Seventeen percent of the sample had been previously identified by their school as requiring special education services. Seventy-six percent of the sample was identified as qualifying for free/reduced lunch, a proxy for low socioeconomic status. All students in the sample were Hispanic. There were six schools and 27 classrooms represented in the sample. Classrooms had, on average, six students represented, with six classrooms being represented by only one student.

Procedures

Recruitment involved obtaining permission from school districts to contact principals at a number of middle schools (in the context of the larger parent project). Teachers in grades 6 and 7 were then briefed about the study, provided information, and an opportunity to answer any questions. If interested and willing, informed consent letters were sent home to students' families. All data for this project was collected as part of a larger effort in three school districts. All examiners were trained by experienced assessment coordinators. All procedures were approved by the Institutional Review Boards of the University of Houston and the University of Texas at Austin.

All assessments were administered by trained, supervised data collectors. Data collectors were hired as part of the TCLD project and included bilingual data collectors to administer Spanish language assessments. Data collectors were trained over a three-week period and training involved formal review of examiner manuals as well as practice with test administration and scoring. Administrative staff members fluent in Spanish were available to assist in examiner training for portions of the assessment protocols conducted in Spanish. All data collectors were tested by project investigators before being approved to test in the schools.

Measures

Four types of measures were obtained from participants: demographic information, objective language tests, self-reported language use, and reading outcomes. We conducted objective assessments of various language constructs and administered a self-report questionnaire evaluating language use preferences across a range of activities and contexts. Objective language assessment included measures (in both Spanish and English) of expressive vocabulary, receptive vocabulary, expressive syntax, and receptive syntax. We conducted standardized reading assessments of word reading accuracy, reading fluency, and reading comprehension.

Demographics. Information about age, gender, socioeconomic status, ethnicity, and racial background of the participants was obtained through the schools.

Language Measures. Students were given assessments of semantics (both receptive and expressive) and syntax (both receptive and expressive) in both Spanish and English. See Chapter 1 (pp. 23-25) for a description of these measures.

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. Reliability for this measure was weak in our sample (alpha = .67). We considered this measure continuously in one set of models in order to evaluate whether a higher level of English usage predicts reading outcomes. In order to also evaluate whether a higher degree of balanced language usage predicts reading, we also considered this measure categorically. Specifically, we inspected the distributions of responses on the measure and divided students into three groups: mostly English usage, mostly Spanish usage, and balanced usage. Additionally, in Chapter 1 we noted that one item from this measure assesses perceived relative language proficiency rather than language usage; specifically, the item asks students to rate which language they are stronger in: Mostly Spanish, Half Spanish/Half English, or Mostly English. Therefore, in order to evaluate whether perceived proficiency is differentially related to reading outcomes compared to ratings of usage, we also considered this single item in our predictive models.

Reading Measures. Participants were individually administered standardized assessments of word reading accuracy, reading fluency, and reading comprehension. Subtests from the Kaufman Test of Educational Achievement-Third Edition (KTEA-3, Kaufman,

2014) were used, including Letter and Word Recognition and Word Recognition Fluency. Split-half reliability estimates for 7th graders taking these subtests are .96 and .89, respectively (Kaufman, 2014). The Gates MacGinite Reading Comprehension Test-Revised (MacGinitie, MacGinitie, Maria, & Dreyer, 2000) was used to assess reading comprehension. Alternate form reliability for the GMTR ranges from .80 to .87 (MacGinitie et al., 2000). Reliabilities for these measures in our sample were adequate and are provided in Table 5. Age-based standard scores were used for the three reading outcomes in the analyses.

Analyses

Descriptive statistics and correlations among the three reading outcomes are shown in Table 5, as well as correlations between reading outcomes, English and Spanish proficiency factor scores, the continuous metric of balance and proficiency, and the self-report measure. A preliminary analysis also considered differences in reading outcomes across the three latent profiles using one-way ANOVA. For descriptive statistics regarding the nine objective language measures used to obtain the English and Spanish factor scores, see Table 1.

Regression diagnostics demonstrated adequate normality, linearity, homogeneity of variance, and independence. Collinearity among the factor scores and continuous metric integrating balance and proficiency was noted, which was not surprising since the continuous metric was computed using the proficiency scores. Thus, separate sets of models were considered in order to evaluate the role of balance. Specifically, one set of models included the continuous metric without proficiency scores, and another set included the latent profiles with proficiency scores.

The relationships between demographics and reading outcomes were examined to determine their potential as covariates. Potential demographic variables included age, gender, free/reduced lunch status (a proxy for socioeconomic status), and special education eligibility. One-way ANOVAs revealed significant effects of special education eligibility (SPED) on word reading ($p < .001$), fluency ($p < .001$), and comprehension ($p < .001$), so this was maintained as a covariate in the analyses. Age was negatively related to word reading ($p = .002$) and fluency ($p < .001$) and was thus maintained as a covariate for these outcomes. Gender and free/reduced lunch status were not related to reading outcomes.

Three sets of hierarchical ordinary least squares regression models were used to evaluate the hypotheses, with one set for each reading outcome (word reading accuracy, oral reading fluency, and reading comprehension). Each step considered the contribution of additional predictors, starting with demographic covariates. In each set of models, Step 1 included demographics. Additionally, for the prediction of reading comprehension, word reading was included as a covariate in Step 1.

Step 2 addressed Hypothesis 1, which predicts that both English and Spanish language skills will differentially predict English reading outcomes, with English proficiency demonstrating a stronger contribution. The English and Spanish proficiency scores were entered in this step.

Step 3 addressed Hypothesis 2, which predicts that metrics integrating balance and proficiency will predict reading outcomes in the context of language proficiency and demographics. Step 3a entered the continuous metric of balance and proficiency. In Step 3b, the continuous metric was removed due to collinearity between the proficiency scores and the balance metric, and the three latent profiles (dummy-coded) were added as another metric that integrates balance and proficiency.

Step 4 addressed Hypotheses 3, which predicts that a self-report measure of language will be related to reading, by entering the self-report measure. Specifically, Step 4a considered the full measure continuously, Step 4b considered the full measure categorically (in order to evaluate whether a balanced level of usage predicts reading), and Step 4c considered item 1 of the scale, which measures perceived language proficiency. Squared semi-partial correlations were computed as estimates of effect size.

Primary analyses were conducted in SAS. Spanish and English factors were exported from MPlus. In order to evaluate whether results with these outputted factor scores were consistent with the latent variables in MPlus, we also conducted regressions in MPlus. Since these results were convergent, we only report the SAS results here. The three latent profiles were exported from MPlus to SAS. More information about how the factor scores and latent profiles were computed can be found in Chapter 1 (pp. 26-29). Once in SAS, the latent profiles were dummy coded for interpretation in regression models such that the coefficients for Profile 2 and Profile 3 were relative to Profile 1.

Because twenty-seven classrooms were represented in our sample, we also computed intraclass correlations to examine the extent to which reading outcomes varied at the classroom level. High intraclass correlations would suggest that regressions should consider the multilevel nature of the data. Intraclass correlations for word reading, reading fluency, and reading comprehension were .02, .05, and .13, respectively, indicating that differences across classrooms accounted for 2%, 5%, and 13% of the variance in these outcomes, respectively. A general rule of thumb is that if the intraclass correlation is 5% or higher, the multilevel nature of the data should be considered (Hox & Roberts, 2011). Therefore, reading fluency and reading comprehension models were conducted using both proc reg and proc mixed programs in SAS. Because both programs produced the same pattern of results for these outcomes, however, we only report results from the proc reg models.

Results

Examination of descriptive statistics for the three reading outcomes (Table 5) indicated that students in this sample fell at least one standard deviation below average (for normative age expectations) across all three reading skills. Lowest scores, on average, were for reading comprehension ($M = 80.51$, $SD = 9.64$), followed by reading fluency ($M = 82.35$, $SD = 12.33$), and word reading ($M = 85.32$, $SD = 13.89$).

Correlations between English language proficiency and reading outcomes were moderate and significant, ranging from $r = .35$ ($p < .001$) for word reading and fluency to $r = .43$ ($p < .001$) for reading comprehension. In contrast, Spanish language proficiency demonstrated a significant but modest correlation with reading comprehension ($r = .16$, $p = .048$), but did not have significant relationships with word reading ($r = .13$, $p =.100$) or reading fluency ($r = .06$, $p = .467$). The continuous metric integrating balance and proficiency was moderately correlated with word reading ($r = .30$, $p < .001$) and reading comprehension ($r = .39$, $p < .001$) but only modestly correlated with reading fluency ($r = .23$, $p = .004$). Notably, the continuous metric was strongly correlated with both English ($r = .67$, $p < .001$) and Spanish ($r = .71$, $p < .001$) proficiency scores, which was not surprising since the continuous metric was computed using these scores. Correlations between self-report metrics (both the full measure and the relative perceived proficiency item) and reading outcomes were generally small and non-significant, with the exception of relative perceived proficiency, which demonstrated a modest but significant correlation with reading fluency ($r = .17$, $p = .031$).

Additionally, one-way ANOVAs were used to inspect group differences in reading outcomes across the three latent profiles. Levene's test was non-significant, verifying the homogeneity of variance assumption ($F = 0.20$, $p = .820$). Figure 2 shows the patterns of word reading, fluency, and comprehension scores across each profile. Profile 1 demonstrated the highest reading scores, followed by Profile 2, followed by Profile 3. These differences across the three profiles were statistically significant for both word reading ($F = 5.98$, $p = .003$) and comprehension ($F = 6.98$, $p = .001$), but were not significant for fluency ($F = 2.62$, $p = .076$). Profile 1, which was characterized by the highest English and Spanish scores as well as the highest degree of balance, demonstrated significantly higher word reading than Profile 2 ($p = .005$) and Profile 3 ($p = .020$), but Profiles 2 and 3 did not differ from one another on word reading ($p = .819$). Profile 1 also demonstrated significantly higher reading comprehension than Profile 2 ($p = .002$) and Profile 3 ($p = .011$), though Profiles 2 and 3 did not differ from one another on comprehension ($p = .795$). Profile 2, which was characterized by the lowest English scores and the second highest Spanish scores, as well as a moderate degree of imbalance, did not significantly differ from Profile 3 on any reading outcome. Profile 3, which was characterized by the lowest Spanish scores and the second highest English scores, as well as a significant imbalance between languages, demonstrated the poorest reading outcomes, though their reading scores were not significantly lower than those of Profile 2.

The Roles of English and Spanish Proficiency in Reading Outcomes (Hypothesis 1)

Results from each set of regression analyses are presented in Tables 6 through 8. Multiple hierarchical regression models tested differential contributions of English and Spanish language proficiency to reading skills. For word reading, consideration of demographics (SPED eligibility and age) accounted for 17% of the variance in Step 1. When English and Spanish factor scores were added in Step 2, these predictors accounted for 25% of the variance in word reading, with a significant contribution from English proficiency (β = 0.32, $p < .001$) but not Spanish proficiency (β = 0.06, p = 0.398). English proficiency demonstrated a unique effect of 9.2% as demonstrated by the squared semi-partial correlation. For reading fluency, consideration of demographics (SPED eligibility and age) accounted for 18% of the variance in Step 1. In Step 2 with the addition of English and Spanish proficiency scores, 27% of the variance in fluency was accounted for, with a significant contribution from English proficiency (β = 0.32, $p < 0.001$) but not Spanish proficiency ($\beta < -0.01$, p = 0.983). English proficiency account for 9.7% of the unique variance in reading fluency. For reading comprehension, Step 1 included SPED eligibility and word reading and accounted for 16% of the variability in comprehension. Word reading contributed 5.2% of the unique variance in comprehension as indicated by the squared semi-partial correlation. In Step 2, 25% of the variability in comprehension was accounted for, with a significant contribution from English proficiency (β = 0.33, $p < 0.001$) but not Spanish proficiency (β = 0.05, p = 0.505). The unique effect of English proficiency on comprehension, estimated through the squared semi-partial correlation, was 9.5%. Word reading was no longer significant in this model (β = 0.15, p = 0.060).

The Role of Balance in Reading Outcomes (Hypothesis 2)

The next step in these regression models considered metrics that integrate both proficiency and balance in addition to proficiency scores and covariates. For word reading, Step 3a added the continuous metric; however, this variable demonstrated collinearity with other predictors in the model (variance inflation factor = 11.57 and tolerance = .08), likely due to the fact that the continuous metric is computed using the two proficiency scores. Thus, in Step 3b, the Spanish proficiency score was removed from the model in order to evaluate the role of balance while also considering English proficiency. Collinearity was not an issue in this model. In this step, English proficiency was a significant predictor of word reading ($\beta = 0.28$, $p = 0.003$), but the continuous metric was not significant ($\beta = 0.05$, $p = 0.576$). The model accounted for 25% of the variance in word reading. In Step 3c, we considered how the latent profiles related to reading skills. We removed the continuous metric of balance and proficiency from this model but included the English and Spanish proficiency scores, as they were not collinear with the dummy-coded latent profiles. This model accounted for 24% of the variance in word reading, with significant contributions from English proficiency ($\beta = 0.30$, $p = 0.003$), but not from the latent profiles or Spanish proficiency.

For reading fluency, Step 3a added the continuous metric, which again demonstrated collinearity with the proficiency scores. When Spanish proficiency was removed from the model in Step 3b, English proficiency was a significant predictor of reading fluency ($\beta = 0.37$, $p < 0.001$), but the continuous metric was not significant ($\beta = -0.08$, $p = 0.432$). The model accounted for 27% of the variance in fluency. When the latent profiles were added in Step 3c (continuous metric removed and proficiency scores re-entered), the model accounted for 27% of the variance in fluency, with a significant contribution from English proficiency ($\beta = 0.39$, $p < 0.001$) but not from the latent profiles or Spanish proficiency.

For reading comprehension, Step 3a added the continuous metric, which was collinear with other predictors in the model. When the Spanish proficiency score was removed in Step 3b, English proficiency was significantly predictive of reading comprehension ($\beta = 0.28$, $p = 0.004$) but the continuous metric was not ($\beta = 0.08$, $p = 0.394$).

This model (covariates, English proficiency, continuous metric) accounted for 25% of the variance in reading comprehension. When the latent profiles were added (continuous metric removed) in Step 3c, 24% of the variance was explained and English proficiency ($\beta = 0.40$, $p < 0.001$) and special education eligibility demonstrated significant contributions. Neither the latent profiles nor Spanish proficiency made significant contributions to comprehension.

The Role of Self-Reported Language Usage and Proficiency in Reading Skills (Hypothesis 3)

In the final step of the regression models, the self-report metric was added in order to evaluate the extent to which this measure contributed additional variance to the prediction of reading skills. The latent profiles were utilized in these analyses so that proficiency scores could remain in the models without problems of multicollinearity. We note that we also ran additional models, one using the continuous metric of balance and proficiency, and another using the proficiency factor scores; these results are reported in text but not in the tables.

For word reading, Step 4a added the continuous self-report measure. This model explained 25% of the variability in word reading but the self-report measure did not make a significant contribution ($\beta = 0.15$, $p = 0.095$). We also considered whether results changed if we considered the continuous metric of balance and proficiency instead of the latent profiles. In this model, the self-report measure was significant ($\beta = 0.20$, $p = 0.006$), and demonstrated a unique contribution of 3.7%. However, in an additional model we removed the continuous metric and added only the English and Spanish proficiency scores, and in that model the self-report measure was not significant ($\beta = 0.15$, $p = 0.088$). These findings suggest that the self-report measure does not significantly predict word reading when English proficiency is also considered in the model.

In Step 4b for word reading, the dummy-coded categories created using cut-offs from the self-report measure were entered in order to evaluate the possible role of balanced usage in reading skills (the continuous self-report measure was removed). This model accounted for 25% of the variance in word reading but neither of the dummy-coded self-report categories made a significant contribution, suggesting that balanced language usage does not contribute to word reading skill. Finally, in Step 4c, the item about perceived relative language proficiency was included as the self-report measure (the categorical self-report variables were removed). This model accounted for 27% of the variability in word reading, and perceived relative proficiency made a significant contribution ($\beta = 0.20$, $p = 0.016$), demonstrating a significant unique contribution of 2.7%, such that a higher level of perceived English proficiency relative to Spanish was associated with better word reading.

For reading fluency, Step 4a which added the continuous self-report measure accounted for 28% of the variability in fluency but the self-report measure did not contribute significantly ($\beta = 0.16$, $p = 0.060$). We again considered whether results changed if we considered the continuous metric of balance and proficiency instead of the latent profiles. In this model, the self-report measure was significant ($\beta = 0.24$, $p < 0.001$) but only demonstrated a unique contribution of 5.3%. In an additional model we removed the continuous metric and added only the English and Spanish proficiency scores, and in that model the self-report measure was also significant ($\beta = 0.17$, $p = 0.048$), demonstrating a X% unique contribution to the prediction of reading fluency. These findings provide mixed support for the role of self-reported language usage in reading fluency. Specifically, results suggest that the self-report measure does not predict fluency when multiple metrics of objective language skill are included in the model (i.e., both proficiency scores and the latent profiles which integrate both proficiency and balance), but self-report does predict fluency when only proficiency scores are included in the model such that a higher level of English usage is related to higher fluency scores.

In Step 4b for reading fluency, the dummy-coded self-report categories were added (continuous self-report measure removed) and this model accounted for 29% of the variance, but neither self-report category made a significant contribution. Finally, when perceived relative language proficiency was added in Step 4c, the model accounted for 29% of the variance in fluency and perceived relative proficiency did contribute significantly ($\beta = 0.20$, $p = 0.014$) and demonstrated unique variance of 2.7%, such that a higher level of perceived English proficiency relative to Spanish was related to higher fluency scores.

For reading comprehension, Step 4a included the continuous self-report measure and accounted for 25% of the variance and the self-report measure did not contribute significantly ($\beta = 0.14$, $p = 0.124$). We again considered whether results changed if we considered the continuous metric of balance and proficiency instead of the latent profiles. In this model, the self-report measure was significant ($\beta = 0.31$, $p < 0.001$) and demonstrated a unique contribution of 3.8%. However, in an additional model we removed the continuous metric and added only the English and Spanish proficiency scores, and in that model the self-report measure was not significant ($\beta = 0.14$, $p = 0.124$).

In Step 4b for reading comprehension, the dummy-coded self-report categories were added and this model accounted for 28% of the variance in comprehension, with a significant unique contribution from the self-report measure of 4.3% such that reporting a higher level of English usage relative to a balanced level of usage was associated with higher comprehension scores ($\beta = 0.24$, $p = 0.003$). Taken together, findings for the role of self-reported language usage in reading comprehension are mixed. Specifically, results suggest that a higher level of English usage relative to Spanish usage is not predictive of comprehension when English proficiency is in the model, but using more English relative to balanced usage (i.e., when the measure was considered categorically) is associated with higher comprehension, even with English proficiency in the model. Finally, when perceived relative language proficiency was added in Step 4c (categorical self-report variables removed), the model accounted for 24% of the variance and relative perceived proficiency did not contribute significantly ($\beta = 0.10$, $p = 0.244$).

A summary of Chapter 2 results is provided in Table 9.

Discussion

The purpose of the present study was to evaluate the roles of L1 and L2 proficiency as well as balance in L2 reading outcomes in an at-risk sample of middle school Spanish-English speaking ELs with reading difficulties. Results demonstrate significant contributions of L2 (English) proficiency to all reading outcomes, while variables that integrate balance and proficiency did not account for additional variance in reading. Additionally, findings were mixed with regard to the roles of student ratings of language usage and proficiency in reading.

The Roles of English and Spanish Language Proficiency in Reading Outcomes

While prior work is mixed with regard to the role of Spanish oral language skills in English reading outcomes, results of the current study clearly indicate that in this population, English oral language skills supplant contributions from Spanish oral language skills in the prediction of English word reading, fluency, and comprehension. Important differences in three specific sample characteristics across studies likely impact differences found among them, as well as conclusions regarding the role of L1 language processes in L2 reading throughout development. First, it is possible that Spanish has an impact on English reading skills, but only at younger ages. Most studies finding support for the role of Spanish language skills in English reading outcomes were conducted at the preschool, kindergarten, or early elementary levels (i.e., Miller et al., 2006; Proctor et al., 2006; Proctor, Harring, & Silverman, 2017), though some studies even at this younger age level did not find support for the role of Spanish skills in English pre-literacy outcomes (i.e., Lonigan et al., 2018). Of course, it is unknown whether Spanish effects would have been evidenced in the present sample, had they been tested several years prior. However, the present negative results raise the possibility that these effects wash out by middle school level. Longitudinal studies that follow students beyond elementary school and into middle school are needed to better understand the impact of development on the relationship between L1 language skills and L2 reading.

A second but related difference across studies is variability with regard to early instructional programming and specifically the language of instruction, which may also have a larger impact at younger ages such as early elementary school. Many of the studies reporting a significant effect of Spanish oral language have been conducted with students who are Spanish language dominant and received the majority of instruction in Spanish and were gradually transitioning to English instruction (i.e., Miller et al., 2006; Proctor et al., 2006). In this context, students may be more likely to need to, and simultaneously be able to, harness their Spanish language skills to support English reading development. In contrast, the middle school students in our sample were receiving all instruction in English and had lower Spanish skills on average relative to their English skills. Even where instructional curricula does have effects, these may dissipate over time. Nakamoto et al. (2008), for instance, found that in sixth grade, Spanish oral language did not predict English reading comprehension in a sample of bilinguals who had all participated in an early transitional bilingual curriculum starting in kindergarten and who had continued to receive some degree of instruction in Spanish throughout elementary school. Unfortunately, we did not have information about our students' early instructional programming, though such information would be important to better understand our findings in the context of prior work.

Third, it is possible that a certain level of Spanish language skill, and/or a certain threshold of English decoding skill, is needed in order for Spanish language to bolster English reading comprehension. Our sample was a high-risk population, with overall low levels of Spanish language and English decoding, and only English language made a contribution to English reading, so these students may have been below such a threshold. Further support for this possibility comes from studies that have found cross-linguistic interactions between Spanish oral language and English decoding (Proctor et al., 2006; Nakamoto et al. 2008) such that Spanish skills contributed to English reading comprehension in strong decoders but not in weaker decoders. Though not the focus of our study since our aim was to better understand these relationships among ELs who are struggling readers, future work should replicate and extend our findings with middle school samples that have a greater representation of average or high Spanish proficiency as well as stronger English decoding. Overall, our findings that English language proficiency and English single word reading predicted English reading comprehension is consistent with well-researched theories of reading development, specifically the Simple View (Gough & Tunmer, 1986), providing novel support for this theory in the context of middle school Spanish-English speaking bilinguals who are also struggling readers.

The Roles of Metrics Integrating Balance and Proficiency in Reading Outcomes

Once we positioned the roles of Spanish and English oral language in English reading, we then sought to evaluate the role of balance between Spanish and English proficiency for English reading outcomes. In constructing our hypotheses about the role of balance in reading, we followed the literature in taking a strengths-based perspective, which suggested that if there are advantages afforded by bilingualism, then they may also extend to reading, which is undergirded in part by cognitive abilities such as EF. We note that although there are a few studies that have found support for a bilingual advantage for reading, these considered bilinguals in aggregate and made comparisons between bilingual and monolingual students across tests of early literacy in young children, rather than considering individual differences within a bilingual sample as we did in this study (Bialystok, Luk, & Kwan, 2009; Bialystok, Majumder, & Martin, 2003).

With our individual differences approach to examining how aspects of bilingualism impact reading, we also followed the literature in choosing balance as a variable that may be related to differences in reading outcomes. There is some work in the bilingual advantage literature supporting the construct of balance (i.e., Vega & Fernandez, 2011; Verreyt et al., 2016), but our study expands upon this by considering reading as the outcome of interest, and also by utilizing metrics of balance that integrate L1 and L2 proficiency levels. The latter is particularly important in at-risk populations such as that of the current study because benefits associated with balance are likely impacted by the overall level(s) of language proficiency.

Studies relating balance to reading are more emergent and have only been conducted with younger bilingual samples (Grimm et al., 2019; Lonigan et al., 2018), though these studies have also utilized methods that integrate both balance and proficiency through latent profile analysis. We expanded upon this work by considering the role of balance and language proficiency in an older sample of children with lower levels of language proficiency and reading achievement than prior studies (and also by including a self-report measure of language, discussed further below). While Lonigan et al. (2018) did not find support for the role of balance, as they found that only English oral language predicted English preliteracy skills, Grimm et al. (2019) found that their balanced group outperformed unbalanced groups on reading comprehension in a sample of second graders. As with our findings, however, the balanced group in the Grimm et al. (2019) study also demonstrated the highest overall performance across English and Spanish language tests, which begs the question of whether it was balance, or simply high English language proficiency, that drove their results. Since we considered both language proficiency levels as well metrics that integrate proficiency with balance, and found that only English proficiency remained significant in these models, our findings suggest that balance (at least as we operationalized it) did not play a significant role. Notably, however, the latent profiles were used in a deterministic rather than a probabilistic manner in our regression models, as profile membership was exported from Mplus categorically rather than probabilistically. This limitation could be addressed in future studies through the use of weighting or resampling techniques.

Despite non-significant findings for the role of metrics that integrate balance and proficiency, it is important to note an interesting trend in the pattern of reading outcomes across the three latent profiles. Specifically, Profile 1 outperformed Profile 2, which outperformed Profile 3, though the difference between Profile 2 and Profile 3 was not statistically significant. However, it is interesting that Profile 2 demonstrated significantly lower English language skills but a higher level of balance than Profile 3, since this trend may point to the role of balance conferring an advantage for reading. We also note that the Grimm et al. (2019) study had findings that were quite similar. Specifically, their moderately unbalanced group (average English/low Spanish) and their very unbalanced group (English dominant, very low Spanish) had comparable levels of English skills, yet the moderately unbalanced group outperformed the very unbalanced group (though as with our study, the difference was not significant). Importantly, this sample's Spanish language scores were nearly as low as ours. Thus, it is possible that additional studies with students who have a higher level of language skill may strengthen this trend in our data, as it is possible a certain level of L1 language ability is needed in order to benefit L2 reading, and this may also be true for conferring an advantage related to balance.

The Roles of Self-Report Indices of Language Usage and Proficiency in Reading Outcomes

We also sought to position a self-report measure of language alongside objective measures in the prediction of reading outcomes in order to better understand how contextual information such as usage of L1 and L2 across contexts may influence L2 reading. We were specifically interested in how a higher level of balanced usage, or alternatively a higher level of English usage, may impact reading outcomes. Importantly, we are not aware of any prior studies that have examined such relationships. Although a mixed pattern of results, small effect sizes, and poor psychometric properties of this measure limit our conclusions, our findings offer novel, preliminary insight into these relationships and can inform future work.

We generally did not find support for self-reported language usage (either balanced usage or English usage) in reading outcomes, with the exception of reading comprehension where a higher level of English usage was associated with better performance, though again we note that this effect was small. In additional analyses, we found preliminary evidence for the role of relative language proficiency in word reading and reading fluency such that a higher level of self-reported English proficiency relative to Spanish proficiency was associated with better performance. Although we are not aware of self-report language measures developed for children, as the items used in the current study were not developed specifically for this purpose, future work may wish to adapt measures that have been developed and validated for these purposes with bilingual adults, as there are a few such measures either currently in development or that have recently become available (Anderson, Mak, Chahi, & Bialystok, 2018). In addition to items about L1 and L2 usage, it is also important to evaluate overall level of perceived L1 and L2 skills rather that relative proficiency. The extent to which these items (usage and proficiency) relate to one another is also unclear, particularly among samples of children, as well as how they may differentially relate to objective language measures and external variables such as reading. Obtaining ratings from multiple informants including students, parents, and/or teachers would also provide a more thorough assessment of these constructs.

Limitations and Future Directions

Current findings should be considered in light of a few limitations. First, as noted above, examination of the roles of L1 and L2 proficiency as well as metrics that integrate balance and proficiency in L2 reading may have yielded different results had we considered a less high-risk sample or had such students also represented within our sample. In particular, it is possible that a higher level of language proficiency, especially in L1, as well as stronger foundational L2 reading skills (i.e., decoding) may have allowed L1 skills to bolster L2 reading outcomes, as has been shown in prior work (Proctor et al., 2006; Nakamoto et al. 2008). However, the current findings are extremely important, novel, and informative given the high-risk nature of this understudied population. For the practical purposes of improving L2 reading in this context, our results suggest that language-based reading intervention in L2 should support the growth of L2 word reading, fluency, and comprehension. This is helpful information, as it suggests that intensive, evidence-based reading interventions already available for monolingual English students with reading difficulties at the middle school level may be similarly effective for bilingual students. A clear next step in this work would be to better understand the impact of such interventions in this population over time, and whether they have a similar effect on reading outcomes as they do when implemented with monolinguals.

It is also important to note that although we did not find support for the role of L1 skills in L2 reading, this does not mean that L1 abilities are not valuable. For instance, it is possible that L1 skills may play a role in functional outcomes such as social skills, adaptive abilities, etc. Similarly, one clear limitation of the present work is that we did not employ measures of Spanish reading achievement in order to evaluate contrasting hypotheses about the roles of L1 and L2 processes as well as metrics of balance in L1 reading achievement. However, we expect that our results would have demonstrated a similar pattern had we included measures of Spanish reading, such that Spanish language skills, but not English skills, may have predicted Spanish reading. Despite this limitation, the current results focused on the prediction of English reading outcomes are important due to the clear functional significance of these skills, as they are the focus of public education at this age level and underlie success in other domains of academic achievement as students are expected to "read to learn" across subjects.

Furthermore, although reading was our primary target of interest because there is currently a stronger evidence base from which to inform intervention for reading relative to domain general skills, future studies would benefit from also including EF tasks in their designs in order to understand whether EF advantages are related to balance in this population, and how EF may or may not serve as a mechanism through which language-related variables impact reading performance.

In addition to various sample characteristics noted above which may have impacted our results, it is also possible that different individual difference variables beyond balance may have yielded different results. For example, the extent to which a bilingual individual switches between using their languages has been implicated in executive function outcomes (Crivello et al., 2016; Hernandez, Dapretto, Mazziotta, & Bookheimer, 2001). Switching can be measured directly through experimental tasks (Crivello et al., 2016), through functional neuroimaging (Hernandez et al., 2001), and also with self-report measures of language switching (Anderson, Mak, Chahi, & Bialystok, 2018). Similarly, Gullifer and Titon (2019) recently proposed the concept of language entropy, which takes into account the extent to which individuals may be compartmentalized versus integrated with their self-report L1 and L2 usage within a single context. Findings from these studies are also mixed, however, in that language switching does not consistently relate to executive function performance (Jylkkä et al., 2017).

Beyond language skills, consideration of other factors may have increased the proportion of explained variance in reading for this sample. For instance, a number of theoretical models of reading comprehension have criticized the Simple View for presenting an oversimplified account of reading comprehension, and have integrated cultural, socioemotional, and contextual factors in addition to cognitive and language-related predictors. For instance, the Componential Model of Reading (Aaron, Joshi, Gooden, & Bentum, 2008), includes domains for cognition, psychological processes, and ecological factors. Psychological variables including reading motivation and acculturation may be particularly important to consider in samples of bilingual children, as suggested by a recent study which sought to validate this model in a bilingual context (Li, Koh, Geva, Joshi, & Chen, 2020). Moreover, with regard to increasing the total variance accounted for in our models, it is possible that our models would have improved had we included pre-literacy skills such as phonological awareness and rapid naming, though we did not have this data available for our sample.

Summary

The current study utilized a nuanced, strengths-based approach to examine how aspects of the bilingual experience impact reading by integrating theoretical frameworks from the bilingual and reading literatures in order to investigate the roles of language proficiency, balance, and self-report measures in word reading, fluency, and comprehension. Results highlight the importance of L2 skills in L2 reading outcomes, with limited findings for the roles of L1 skills or balance in reading. Findings for the self-report measure were mixed and should be considered perhaps in different ways for future work. This work sheds light on the ways in which variability in language skills impacts an important academic domain in a highly at-risk and understudied population with the goal of using this enhanced understanding to inform future intervention efforts.

References

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research*, *80*(2), 207-245.

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

Anderson, J. A., Mak, L., Chahi, A. K., & Bialystok, E. (2018). The language and social background questionnaire: Assessing degree of bilingualism in a diverse population. *Behavior Research Methods*, *50*(1), 250-263.

Antón, E., Duñabeitia, J. A., Estévez, A., Hernández, J. A., Castillo, A., Fuentes, L. J., ... & Carreiras, M. (2014). Is there a bilingual advantage in the ANT task? Evidence from children. *Frontiers in Psychology*, *5*, 398.

Archila-Suerte, P., Woods, E. A., Chiarello, C., & Hernandez, A. E. (2018). Neuroanatomical profiles of bilingual children. *Developmental Science*, *21*(5), e12654.

Bialystok, E. (1999). Cognitive complexity and attentional control in the bilingual mind. *Child development*, *70*(3), 636-644.

Bialystok, E., & Barac, R. (2012). Emerging bilingualism: Dissociating advantages for metalinguistic awareness and executive control. *Cognition*, *122*(1), 67-73.

Bialystok, E., Craik, F., & Luk, G. (2008). Cognitive control and lexical access in younger and older bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 859-873.

Bialystok, E., Majumder, S., & Martin, M. M. (2003). Developing phonological awareness: Is there a bilingual advantage?. *Applied Psycholinguistics*, *24*(1), 27-44.

Bialystok, E., & Martin, M. M. (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, *7*(3), 325-339.

Bialystok, E., Luk, G., & Kwan, E. (2005). Bilingualism, biliteracy, and learning to read: Interactions among languages and writing systems. *Scientific studies of reading*, *9*(1), 43-61.

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

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology*, *53*(1), 371-399.

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

Branum-Martin, L., Mehta, P. D., Francis, D. J., Foorman, B. R., Cirino, P. T., Miller, J. F., & Iglesias, A. (2009). Pictures and words: Spanish and English vocabulary in classrooms. *Journal of Educational Psychology*, *101*(4), 897-911.

Butterfuss, R., & Kendeou, P. (2018). The role of executive functions in reading comprehension. *Educational Psychology Review*, *30*(3), 801-826.

Carlson, S. M., & Meltzoff, A. N. (2008). Bilingual experience and executive functioning in young children. *Developmental Science*, *11*(2), 282-298.

Catts, H. W., Hogan, T. P., Adlof, S. M., & Barth, A. E. (2003, June). The simple view of reading changes over time. In *annual meeting of the Society for Scientific Study of Reading, Boulder, CO*.

Chomsky, N., & Miller, G. A. (1958). Finite state languages. *Information and control*, *1*(2), 91-112.

Cirino, P. T., Miciak, J., Ahmed, Y., Barnes, M. A., Taylor, W. P., & Gerst, E. H. (2019). Executive function: Association with multiple reading skills. *Reading and writing*, *32*(7), 1819-1846.

Crivello, C., Kuzyk, O., Rodrigues, M., Friend, M., Zesiger, P., & Poulin-Dubois, D. (2016). The effects of bilingual growth on toddlers' executive function. *Journal of Experimental Child Psychology*, *141*, 121-132.

De Feyter, J. J., & Winsler, A. (2009). The early developmental competencies and school readiness of low-income, immigrant children: Influences of generation, race/ethnicity, and national origins. *Early Childhood Research Quarterly*, *24*(4), 411-431.

De Renzi, A., & Vignolo, L. A. (1962). Token test: A sensitive test to detect receptive disturbances in aphasics. *Brain: A Journal of Neurology*, *85*, 665-678.

Farnia, F., & Geva, E. (2013). Growth and predictors of change in English language learners' reading comprehension. *Journal of Research in Reading*, *36*(4), 389-421.

Fletcher, J. M., Denton, C., & Francis, D. J. (2005). Validity of alternative approaches for the identification of learning disabilities: Operationalizing unexpected underachievement. *Journal of Learning Disabilities*, *38*(6), 545-552.

Follmer, D. J. (2018). Executive function and reading comprehension: A meta-analytic review. *Educational Psychologist*, *53*(1), 42-60.

Foorman, B. R., Koon, S., Petscher, Y., Mitchell, A., & 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*(3), 884-899.

Francis, D. J., Rivera, M., Lesaux, N., Kieffer, M., & 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*.

Gathercole, V. C. M. (2017). Straw man: Who thought exposure was the ONLY factor?. *Bilingualism: Language and Cognition*, *20*(1), 23-24.

Geva, E., & Massey-Garrison, A. (2013). A comparison of the language skills of ELLs and monolinguals who are poor decoders, poor comprehenders, or normal readers. *Journal of Learning Disabilities*, *46*(5), 387-401.

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (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*(3), 594-615.

Goodman, E. B. J. C. (1997). On the inseparability of grammar and the lexicon: Evidence from acquisition, aphasia and real-time processing. *Language and Cognitive Processes*, *12*(5-6), 507-584.

Gottardo, A., & 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*(2), 330-344.

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

Gray, S., O'Connell, A. A., Restrepo, M. A., Yeomans-Maldonado, G., Bengochea, A., Mesa, C., ... & Mesa, C. (2018). The dimensionality of oral language in kindergarten Spanish–English dual language learners. *Journal of Speech, Language, and Hearing Research*, *61*(11), 2779-2795.

Genesee, F., Lindholm-Leary, K., Saunders, W., & Christian, D. (2005). English language learners in US schools: An overview of research findings. *Journal of Education for Students Placed at Risk*, *10*(4), 363-385.

Grimm, R., Solari, E. J., Gerber, M. M., Nylund-Gibson, K., & Swanson, H. L. (2019). A Cross-Linguistic Examination of Heterogeneous Reading Profiles of Spanish-Speaking Bilingual Students. *The Elementary School Journal*, *120*(1), 109-131.

Hammer, C. S., Jia, G., & 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*(1), 4-9.

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

Hart, S. A., Logan, J. A., Thompson, L., Kovas, Y., McLoughlin, G., & Petrill, S. A. (2016). A latent profile analysis of math achievement, numerosity, and math anxiety in twins. *Journal of educational psychology*, *108*(2), 181-193.

Hernandez, A. E., Dapretto, M., Mazziotta, J., & Bookheimer, S. (2001). Language switching and language representation in Spanish–English bilinguals: An fMRI study. *NeuroImage*, *14*(2), 510-520.

Hernandez, D. J., Denton, N. A., & Blanchard, V. L. (2011). Children in the United States of America: A statistical portrait by race-ethnicity, immigrant origins, and language. *The ANNALS of the American Academy of Political and Social Science*, *633*(1), 102-127.

Hemphill, F. C., & Vanneman, A. (2011). Achievement Gaps: How Hispanic and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2011-459. *National Center for Education Statistics*.

Hindman, A. H., & Wasik, B. A. (2015). Building vocabulary in two languages: An examination of Spanish-speaking dual language learners in Head Start. *Early Childhood Research Quarterly*, *31*, 19-33.

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*(1), 4-14.

Hox, J., & Roberts, J. K. (Eds.). (2011). *Handbook of advanced multilevel analysis*. Psychology Press.

Jylkkä, J., Soveri, A., Wahlström, J., Lehtonen, M., Rodríguez-Fornells, A., & Laine, M. (2017). Relationship between language switching experience and executive functions in bilinguals: an Internet-based study. *Journal of Cognitive Psychology*, *29*(4), 404-419.

Kaufman, A. S. (2014). Kaufman test of educational achievement–third edition (KTEA-3) technical & interpretive manual. *Bloomington, MN: Pearson*.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773-795.

Kassai, R., Futo, J., Demetrovics, Z., & Takacs, Z. K. (2019). A meta-analysis of the experimental evidence on the near-and far-transfer effects among children's executive function skills. *Psychological Bulletin*, *145*(2), 165-188.

Kena, G., Musu-Gillette, L., Robinson, J., Wang, X., Rathbun, A., Zhang, J., ... & Velez, E. D. V. (2015). The Condition of Education 2015. NCES 2015-144. *National Center for Education Statistics*.

Kim, D. H., Lambert, R. G., & Burts, D. C. (2018). Are young dual language learners homogeneous? Identifying subgroups using latent class analysis. *The Journal of Educational Research*, *111*(1), 43-57.

Law, J., Garrett, Z., & Nye, C. (2003). Speech and language therapy interventions for children with primary speech and language delay or disorder. *Cochrane Database of Systematic Reviews*, (3).

Li, M., Koh, P. W., Geva, E., Joshi, R., & Chen, X. (2020). The componential model of reading in bilingual learners. *Journal of Educational Psychology*, *112*(8), 1532–1545.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaire: A web-based interface for bilingual research. *Behavior research methods*, *38*(2), 202-210.

Li, P., Zhang, F., Tsai, E., & 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, C. J., & Milburn, T. F. (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*(8), 2185-2198.

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual review of psychology*, *51*(1), 201-226.

MacGinitie, W. H., MacGinitie, R. K., Maria, K., & Dreyer, L. G. (2000). Gates–MacGinitie Reading Tests (4th ed.). Itasca, IL: Riverside.

Marian, V., Blumenfeld, H. K., & 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*(4), 940-967.

Marsh, H. W., Lüdtke, O., Trautwein, U., & Morin, A. J. (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*(2), 191-225.

Martin, N. A., & Brownell, R. (2010). Receptive One-Word Picture Vocabulary Test Fourth Edition (ROWPVT-4) Spanish/Bilingual Version. *Novato: Academic Therapy Publications*.

Meeusen, C., Meuleman, B., Abts, K., & Bergh, R. (2018). Comparing a variable-centered and a person-centered approach to the structure of prejudice. *Social Psychological and Personality Science*, *9*(6), 645-655.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of “far transfer” evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512-534.

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. *Learning Disabilities Research & Practice*, *21*(1), 30-43.

Morales, J., Gómez-Ariza, C. J., & Bajo, M. T. (2013). Dual mechanisms of cognitive control in bilinguals and monolinguals. *Journal of Cognitive Psychology*, *25*(5), 531-546.

Morris, R. D., Fletcher, J. M., & Francis, D. J. (1992). Conceptual and psychometric issues in the neuropsychologic assessment of children: Measurement of ability discrepancy and change. *Handbook of neuropsychology*, *6*, 341-341.

Mueller, R. O., & Hancock, G. R. (2008). Best practices in structural equation modeling. *Best practices in quantitative methods*, *488-508*.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005). Batería III: Woodcock- Muñoz: Pruebas de Aprovechamiento. Itasca, IL: Riverside Publishing.

Muthén, L. K., & Muthén, B. O. (2012). Mplus statistical modeling software: Release 7.0. *Los Angeles, CA: Muthén & Muthén*.

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

Nylund, K. L., Asparouhov, T., & Muthén, B. O. (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*(4), 535-569.

Ortman, J. M., & Shin, H. B. (2011, August). Language projections: 2010 to 2020. In *American Sociological Association Annual Meeting*.

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive psychology*, *66*(2), 232-258.

Paap, K. R., & Sawi, O. (2014). Bilingual advantages in executive functioning: problems in convergent validity, discriminant validity, and the identification of the theoretical constructs. *Frontiers in psychology*, *5*, 962.

Passel, J. S., Cohn, D. V., & Lopez, M. H. (2011). Hispanics account for more than half of nation's growth in past decade. *Washington, DC: Pew Hispanic Center*, 1-7.

Pérez-Leroux, A. T., Castilla-Earls, A. P., & 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(*2), 313-327.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific studies of Reading*, *18*(1), 22-37.

Pinker, S. (1998). Words and rules. *Lingua*, *106*(1), 219-242.

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

Proctor, C. P., Harring, J. R., & Silverman, R. D. (2017). Linguistic interdependence between Spanish language and English language and reading: A longitudinal exploration from second through fifth grade. *Bilingual Research Journal*, *40*(4), 372-391.

Rosselli, M., Ardila, A., Lalwani, L. N., & Vélez-Uribe, I. (2016). The effect of language proficiency on executive functions in balanced and unbalanced Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *19*(3), 489-503.

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

Saunders, W. M., & Marcelletti, D. J. (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*(2), 139-156.

Semel, E. M., Wiig, E. H., Secord, W., & Langdon, H. W. (2006). *CELF 4: Clinical Evaluation of Language Fundamentals 4: Spanish Edition*. PsychCorp.

Scammacca, N. K., Roberts, G. J., Cho, E., Williams, K. J., Roberts, G., Vaughn, S. R., & Carroll, M. (2016). A century of progress: Reading interventions for students in grades 4–12, 1914–2014. *Review of Educational Research*, *86*(3), 756-800.

Schumacker, R. E., & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling*. Psychology Press.

Sheng, L., Lu, Y., & Gollan, T. H. (2014). Assessing language dominance in Mandarin–English bilinguals: Convergence and divergence between subjective and objective measures. *Bilingualism: Language and Cognition*, *17*(2), 364-383.

Snijders, T. A., & Bosker, R. J. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Swanson, H. L., Orosco, M. J., Lussier, C. M., Gerber, M. M., & Guzman-Orth, D. A. (2011). The influence of working memory and phonological processing on English language learner children's bilingual reading and language acquisition. *Journal of Educational Psychology*, *103*(4), 838-856.

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1193-1208.

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition*, *22*(3), 516-536.

Van den Broek, P., Young, M., Tzeng, Y., & Linderholm, T. (1999). The landscape model of reading: Inferences and the online construction of a memory representation. *The construction of mental representations during reading*, 71-98.

Vaughn, K. A., & Hernandez, A. E. (2018). Becoming a balanced, proficient bilingual: Predictions from age of acquisition & genetic background. *Journal of neurolinguistics*, *46*, 69-77.

Vega, C., & Fernandez, M. (2011). Errors on the WCST correlate with language proficiency scores in Spanish-English bilingual children. *Archives of Clinical Neuropsychology*, *26*(2), 158-164.

Vellutino, F. R., Fletcher, J. M., Snowling, M. J., & Scanlon, D. M. (2004). Specific reading disability (dyslexia): What have we learned in the past four decades? *Journal of Child Psychology and Psychiatry*, *45*(1), 2-40.

Verreyt, N., Woumans, E. V. Y., Vandelanotte, D., Szmalec, A., & Duyck, W. (2016). The influence of language-switching experience on the bilingual executive control advantage. *Bilingualism: Language and Cognition*, *19*(1), 181-190.

White, L. J., & Greenfield, D. B. (2017). Executive functioning in Spanish-and English-speaking Head Start preschoolers. *Developmental Science*, *20*(1), e12502.

Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical evaluation of language fundamentals: CELF-5*. Pearson.

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. A. (2007). Woodcock-Johnson normative update tests of cognitive abilities. *Rolling Meadows, IL: Riverside*.

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

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111-125.

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	.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	χ2	*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	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
6	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

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 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.

*** $p < .001$

Table 4. Summary of Chapter 1 Results

	Hypothesis/Aim	Result
1	Confirmatory models differentiating between language (English vs. Spanish) and subskill (semantics vs. syntax) will provide the best fit to the data.	A two-factor model differentiating between English and Spanish language processes provided the best fit to the data.
2	Latent profile analysis will demonstrate four subgroupings of students in the sample, characterized by differences in L1 and L2 proficiency and balance.	Three latent profiles emerged in the sample, which were characterized by differences in L1 and L2 proficiency and balance.
3	Latent profiles will differentiate according to L1 and L2 factor proficiency scores obtained through the factor model.	Latent profiles differed from one another on L1 and L2 proficiency scores in expected directions.
4	Latent profiles will differ on a continuous metric that integrates proficiency and balance.	Latent profiles differed from one another on the continuous metric in expected directions.
5	Latent profiles will differ on a self-report measure of language usage.	Latent profiles differed from one another on a self-report measure of language usage in expected directions, though not all pairwise comparisons were significant.
6	We will evaluate a more parsimonious approach to the measurement of language.	Results using English and Spanish expressive *Picture Vocabulary* measures as well as results using English and Spanish *Memory for* Sentences (expressive syntax) provided the same general pattern of results as the full battery of nine language measures.

Table 5. Descriptive Statistics, Reliabilities, and Correlations among Language and Reading Variables

		1	2	3	4	5	6	7	8
1	Word Reading	--							
2	Reading Fluency	.73**	--						
3	Reading Comprehension	.34**	.45**	--					
4	English Proficiency	.35**	.35**	.43**	--				
5	Spanish Proficiency	.13	.06	.16*	.04	--			
6	Continuous Metric of Balance and Proficiency	.30**	.23*	.39**	.67**	.71**	--		
7	Self-Report Language Usage	.10	.15	.12	.24*	-.56**	-.23*	--	
8	Self-Reported Perceived Relative Proficiency Item	.13	.17*	.10	.20*	-.53**	-0.25*	.72**	--
	Mean	85.32	82.35	80.51	0	0	0.06	2.17	1.99
	SD	13.89	12.33	9.64	0.83	0.94	1.01	0.35	0.85
	Reliability^{+}	.94	.93	.77				.67	

$^{*}p < .05$

$^{**}p < .001$

$^{+}$ Reliabilities for the nine objective language measures which were used to compute the English and Spanish proficiency scores, as well as the Continuous Metric of Balance and Proficiency, are shown in Table 1.

Table 6. Regression Analyses Predicting Word Reading (*N* = 161)

	B	*SE b*	*β*	*Squared Semi-Partial Correlation*
	Step 1: Covariates			
SPED Eligibility Status+	-12.53**	2.65	-0.34	0.117
Age	-3.99**	1.34	-0.22	0.046
Adjusted R^2	.17**			
	Step 2: Proficiency Scores			
English Proficiency	5.28**	1.19	0.32	0.092
Spanish Proficiency	0.94	1.11	0.06	0.003
Adjusted R^2	.25**			
	Step 3a: Continuous Metric of Balance and Proficiency			
Continuous Metric	-1.58	3.20	-0.12	0.001
Adjusted R^2	.25**			
	Step 3b: Spanish Proficiency Score Removed			
Continuous Metric	0.75	1.34	0.05	0.001
Adjusted R^2	.21**			
	Step 3c: Latent Profiles with Proficiency Scores, Continuous Metric Removed			
Profile 2++	-1.06	3.30	-0.04	<0.001
Profile 3++	-2.37	5.54	-0.06	0.001
Adjusted R^2	.24**			
	Step 4a: Self-Reported Language Usage as a Continuous Variable			
Self-Report Usage	5.87	3.49	0.15	0.014
Adjusted R^2	.25**			
	Step 4b: Self-Reported Language Usage as a Categorical Variable (Continuous Variable Removed)			
English Usage+++	1.64	2.74	0.05	0.002
Spanish Usage+++	-4.62	2.71	-0.13	0.014
Adjusted R^2	.25**			
	Step 4c: Self-Reported Perceived Relative Proficiency (Categorical Variables Removed)			

Self-Report Proficiency	3.33*	1.36	.20	0.027
Adjusted R^2		.27**		

Note: All steps include covariates (SPED eligibility and age). Steps for entering and removing additional variables are based on hypotheses and are described in text. English and Spanish proficiency scores are included in Step 4 models.

*$p < .05$ **$p < .01$

$^{+}$Reference group = Not eligible for special education

$^{++}$Reference group = Profile 1

$^{+++}$Reference group = Balanced usage

Table 7. Regression Analyses Predicting Reading Fluency (*N* = 161)

	B	*SE b*	*β*	*Squared Semi-Partial Correlation*
		Step 1: Covariates		
SPED Eligibility Status[+]	-10.32**	2.33	-0.32	0.101
Age	-4.55**	1.18	-0.28	0.076
Adjusted R^2		.18**		
		Step 2: Proficiency Scores		
English Proficiency	4.82**	1.04	0.32	0.097
Spanish Proficiency	-0.02	0.97	< -0.01	<0.001
Adjusted R^2		.27**		
		Step 3a: Continuous Metric of Balance and Proficiency		
Continuous Metric	-5.09	2.77	-0.42	0.015
Adjusted R^2		.28**		
		Step 3b: Spanish Proficiency Score Removed		
Continuous Metric	-0.92	1.17	-0.08	0.003
Adjusted R^2		.27**		
		Step 3c: Latent Profiles with Proficiency Scores, Continuous Metric Removed		
Profile 2[++]	2.70	2.88	0.11	0.004
Profile 3[++]	-0.74	4.82	-0.02	<0.001
Adjusted R^2		.27**		
		Step 4a: Self-Reported Language Usage as a Continuous Variable		
Self-Report Usage	5.76	3.04	0.16	0.012
Adjusted R^2		.28**		
		Step 4b: Self-Reported Language Usage as a Categorical Variable (Continuous Variable Removed)		
English Usage[+++]	4.11	2.37	0.13	0.013
Spanish Usage[+++]	-3.60	2.35	-0.11	0.010
Adjusted R^2		.29**		
		Step 4c: Self-Reported Perceived Relative Proficiency (Categorical Variables Removed)		
Self-Report Proficiency	2.95*	1.19	0.20	0.027

Adjusted R^2	.29**

Note: All steps include covariates (SPED eligibility and age). Steps for entering and removing additional variables are based on hypotheses and are described in text. English and Spanish proficiency scores are included in Step 4 models.

*$p < .05$ **$p < .01$

+Reference group = Not SPED eligible

++Reference group = Profile 1

+++Reference group = Balanced usage

Table 8. Regression Analyses Predicting Reading Comprehension (*N* = 160)

	B	*SE b*	*β*	*Squared Semi-Partial Correlation*
		Step 1: Covariates		
SPED Eligibility Status+	-6.43**	1.97	-0.25	0.056
Word Reading	0.17**	0.05	0.24	0.052
Adjusted R^2		.16**		
		Step 2: Proficiency Scores		
English Proficiency	3.86**	0.86	0.33	0.095
Spanish Proficiency	0.51	0.76	0.05	0.002
Adjusted R^2		.25**		
		Step 3a: Continuous Metric of Balance and Proficiency		
Continuous Metric	1.30	2.21	0.14	0.002
Adjusted R^2		.24**		
		Step 3b: Spanish Proficiency Score Removed		
Continuous Metric	0.79	0.93	0.08	0.003
Adjusted R^2		.21**		
		Step 3c: Latent Profiles with Proficiency Scores, Continuous Metric Removed		
Profile 2++	2.16	2.28	0.11	0.004
Profile 3++	1.50	3.85	0.05	0.001
Adjusted R^2		.24**		
		Step 4a: Self-Reported Language Usage as a Continuous Variable		
Self-Report Usage	3.78	2.44	0.14	0.011
Adjusted R^2		.25**		
		Step 4b: Self-Reported Language Usage as a Categorical Variable (Continuous Variable Removed)		
English Usage+++	5.74**	1.87	0.24	0.043
Spanish Usage+++	0.40	1.87	0.02	<0.001
Adjusted R^2		.28**		
		Step 4c: Self-Reported Perceived Relative Proficiency (Categorical Variables Removed)		
Self-Report Proficiency	1.13	0.97	0.10	0.007

Adjusted R^2	.24**

Note: All steps include covariates (SPED eligibility and word reading). Steps for entering and removing additional variables are based on hypotheses and are described in text. English and Spanish proficiency scores are included in Step 4 models.

$*p < .05$ $**p < .01$

$^{+}$Reference group = Not eligible for special education

$^{++}$Reference group = Profile 1

$^{+++}$Reference group = Balanced usage

Table 9. Summary of Chapter 2 Results

	Hypothesis	Result: Word Reading	Result: Fluency	Result: Comprehension
1	English language proficiency will have a larger impact on English reading outcomes compared to Spanish language proficiency.	For all three outcomes, English proficiency made a significant contribution, Spanish proficiency did not.		
2	Objective metrics integrating both balance and proficiency will predict reading outcomes over and above the contribution of language proficiency.	For all three outcomes, a continuous metric of balance and proficiency was not significant. A categorical metric of balance and proficiency (latent profiles) did not predict any reading outcomes when proficiency scores were included in the model.		
3a	A self-report measure of language usage will demonstrate an incremental contribution to reading outcomes in the context of proficiency and balance metrics.	A higher level of self-reported English usage relative to Spanish usage was not significant when English proficiency was also in the model. A higher level of English usage relative to balanced usage was also not significant.	A higher level of self-reported English usage relative to Spanish usage was not significant when all other language predictors were in the model. A higher level of English usage relative to balanced usage was also not significant.	A higher level of self-reported English usage relative to Spanish usage was not significant when English proficiency was also in the model. However, a higher level of English usage relative to balanced usage was significant.
3b	A self-report measure of perceived relative language proficiency will demonstrate an incremental contribution to reading outcomes in the context of proficiency and balance metrics.	A higher level of self-reported English proficiency relative to Spanish proficiency was significant.	A higher level of self-reported English proficiency relative to Spanish proficiency was significant.	A higher level of self-reported English proficiency relative to Spanish proficiency was not significant.

Figure 1. Two-Factor English/Spanish Model

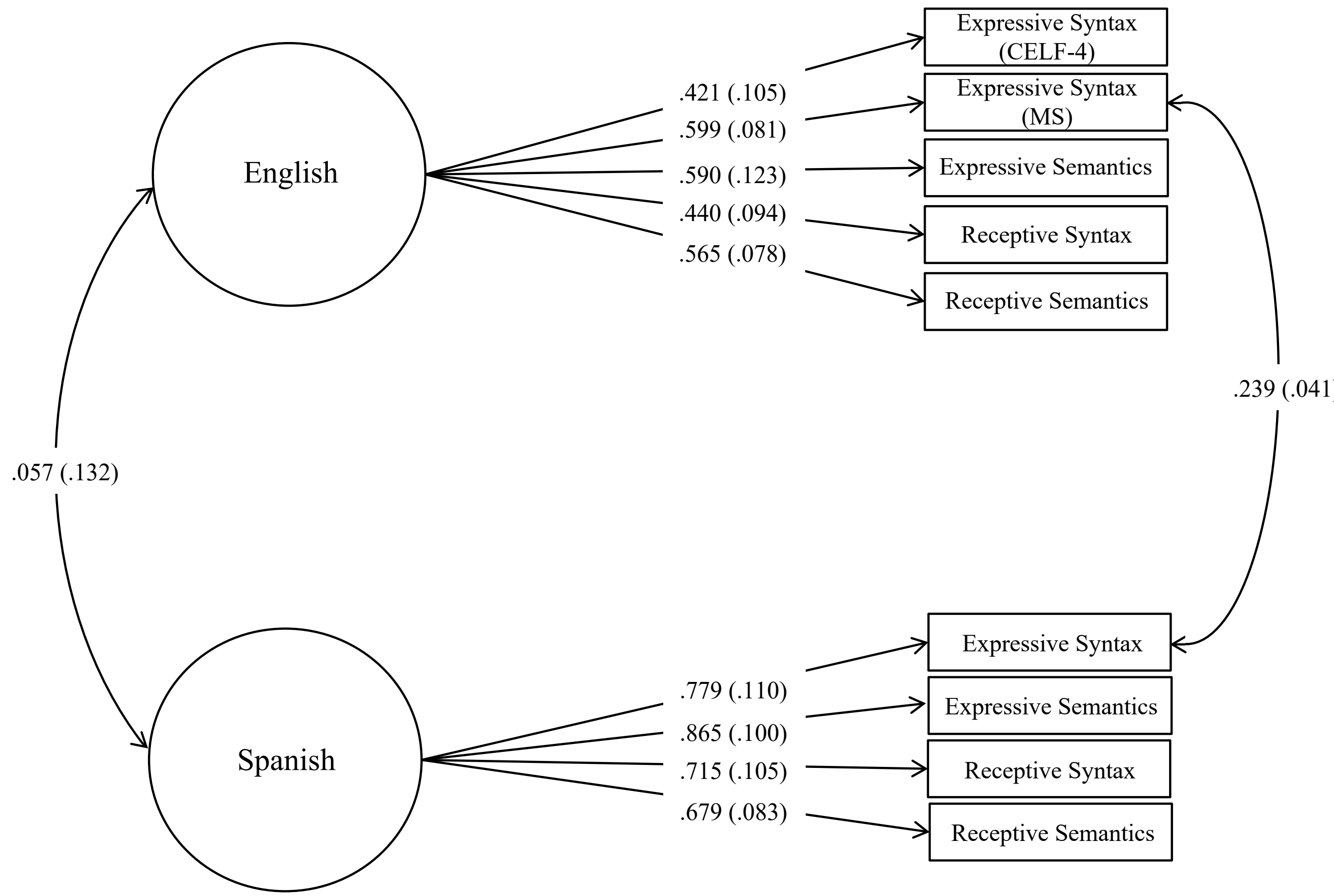
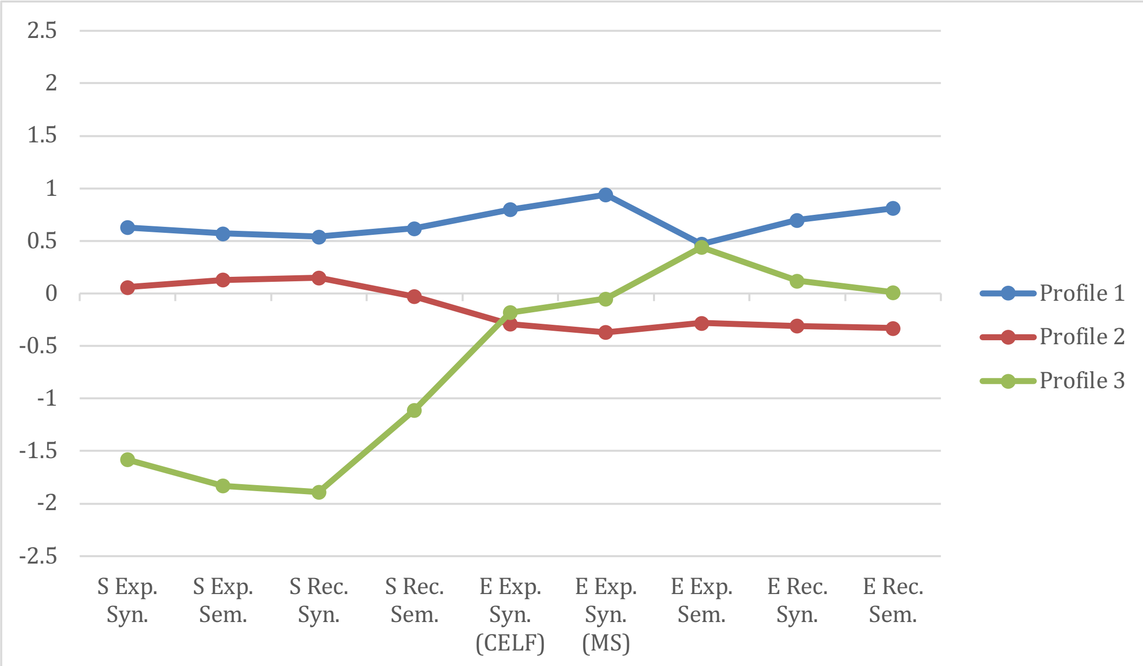
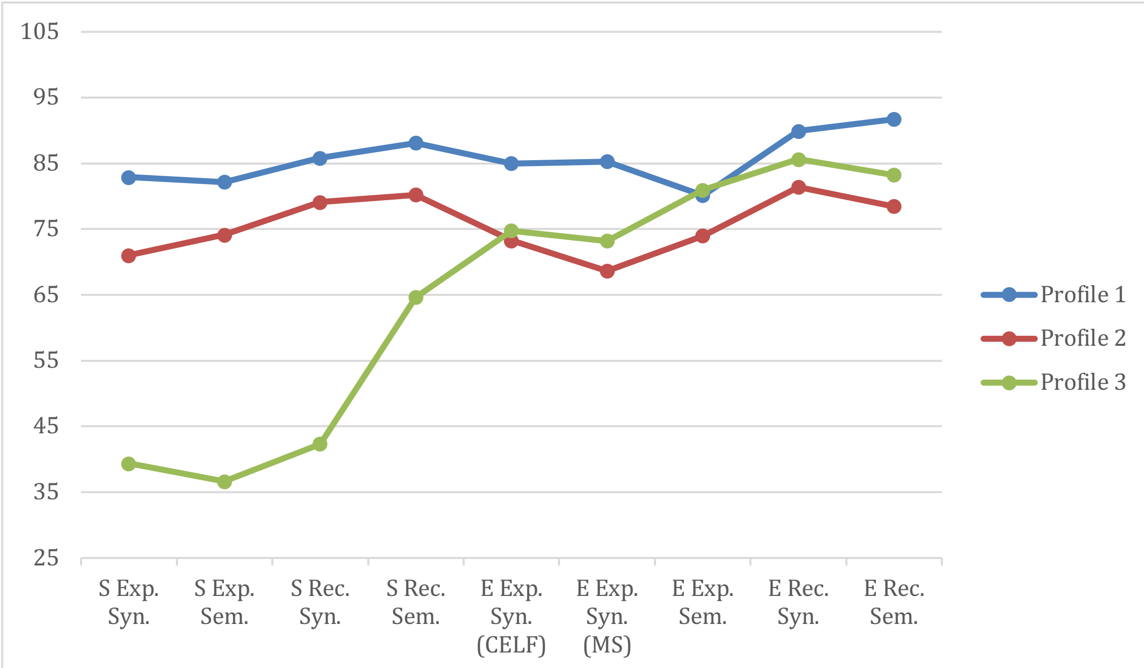


Figure 2. Standardized Raw Score Performance on Language Measures across Latent Profiles



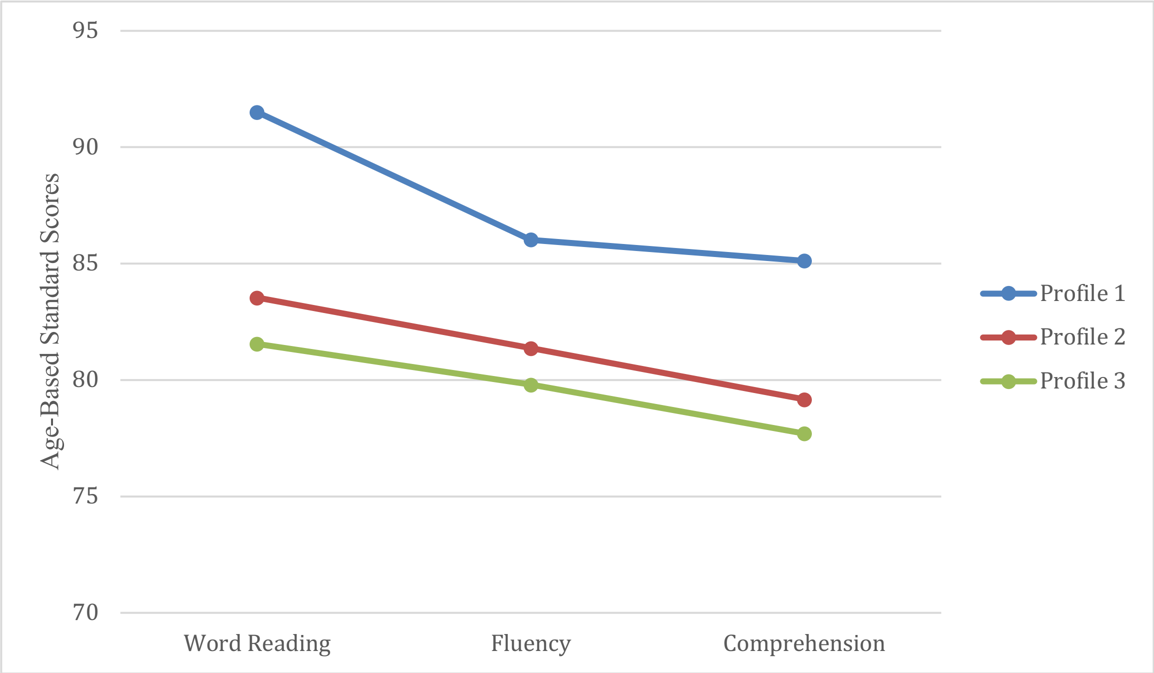
Note. S Exp. Syn. = Spanish Expressive Syntax; S Exp. Sem. = Spanish Expressive Semantics; S Rec. Syn. = Spanish Receptive Syntax; S Rec. Sem. = Spanish Receptive Semantics; E Exp. Syn. (CELF) = English Expressive Syntax measured with the Clinical Evaluation of Language Fundamentals Sentence Assembly 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.

Figure 4. Reading Performance across Latent Profiles



Note. Profile 1 demonstrated significantly higher word reading than Profile 2 (p = .005) and Profile 3 (p = .020), but Profiles 2 and 3 did not differ from one another (p = .819). Profile 1 also demonstrated significantly higher reading comprehension than Profile 2 (p = .002) and Profile 3 (p = .011), though Profiles 2 and 3 did not differ from one another (p = .795). Differences across the three profiles were not significant for fluency (F = 2.62, p = .076).