Reliability and Validity Evidence for the English and Spanish Preschool Narrative Language Measures-Listening

Trina D. Spencer, PhD

Department of Child & Family Studies, University of South Florida, Tampa, FL

Marilyn S. Thompson PhD

Sanford School of Social & Family Dynamics, Arizona State University, Tempe, AZ

Douglas B. Petersen PhD

Department of Communication Disorders, Brigham Young University, Provo, UT

Yixing Liu, PhD

Beijing University of Chinese Medicine, Beijing, China

M. Adelaida Restrepo PhD

College of Health Solutions, Arizona State University, Tempe, AZ

Reference:

Spencer, T. D., Thompson, M. S., Petersen, D. B., Liu, Y., & Restrepo, M. A. (2023). Reliability and validity evidence for the English and Spanish preschool Narrative Language Measures Listening. *Early Childhood Research Quarterly, 64*, 148-161. https://doi.org/10.1016/j.ecresq.2023.02.005

Author Note

Trina D. Spencer https://orcid.org/0000-0002-3531-8276

Marilyn S. Thompson https://orcid.org/0000-0002-4419-6290

Douglas B. Petersen https://orcid.org/0000-0003-2675-4217

Yixing Liu https://orcid.org/0000-0001-5992-8450

M. Adelaida Restrepo https://orcid.org/0000-0001-6210-3759

Dr. Petersen and Dr. Spencer are co-authors of the *CUBED* assessment. They have a financial interest in Language Dynamics Group LLC, which provides free access to the CUBED assessment. Dr. Restrepo is a co-author of the Spanish Preschool Narrative Language Measures Listening, but she does not have a financial relationship with Language Dynamics Group LLC.

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A140093 for $1,481,960 to Northern Arizona University and University of South Florida. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

For young Spanish-speaking children entering U. S. schools, it is imperative that educators foster growth in the home language and in the language of instruction to the fullest extent possible. Monitoring language development over time is crucial for promoting language development because it allows educators to individualize student instruction. The Narrative Language Measures-Listening (NLM-Listening) subtest of the CUBED assessment was designed as a curriculum-based benchmark and progress monitoring tool for eliciting narrative language samples from preschool children in an authentic manner. In both English and Spanish, NLM-Listening includes 25 preschool alternate forms equated on story structure and language complexity. Standardized administration and real time scoring procedures enhance the NLM-Listening's reliability and feasibility. This study offers a rigorous examination of the reliability and validity evidence for using the Spanish and English NLM-Listening with preschool children. Using a Latin-square design to randomize order of administration, 126 three- to four-year-old children attending Head Start preschool received up to 25 forms of the Spanish (*N* = 61) or English (*N* = 65) versions, and two criterion measures of language in Spanish or English, respectively, within a few weeks of each other. Results indicated that the NLM-Listening has strong alternate-form reliability and strong evidence of validity based on unidimensional factor analyses and moderately high correlations with criterion measures of language. The NLM-Listening is a promising tool for monitoring preschoolers' language development in English and Spanish.

Reliability and Validity Evidence for the English and Spanish Preschool Narrative Language Measures-Listening

Being bilingual is an asset and comes with many advantages academically and professionally in the U.S and worldwide. However, many bilingual students in the U.S. perform poorly on national and state summative reading and writing assessments. For example, over 92% of English learners in fourth grade read below a proficient level on the National Assessment of Education Progress (NAEP, 2019). The majority of those students have difficulty with reading comprehension which is highly dependent on oral language comprehension (Kieffer & Vukovic, 2012; Nakamoto et al., 2007). The cluster of oral language skills needed for school success is often referred to as academic language (Schleppegrell, 2001). Specialized in its function, academic language is used to express and acquire knowledge. In its form, academic language comprises complex word-, sentence-, and discourse-level structures that are typical of written language (Phillips Galloway et al., 2019). Students whose home language differs from the mainstream language or whose parents do not standardly use academic language registers are disadvantaged in school because they not only need to gain command of this specialized language to express knowledge, but they have to struggle to acquire knowledge taught through this unfamiliar medium (Gottlieb & Ernst-Slavit, 2014).

Bilingual students, especially those learning English as a second language, and much like many other students in U.S. public schools, need additional exposure and support in acquiring academic language skills so that when such language is presented to them in either oral or written form, they are prepared to comprehend and use it (Cirino et al., 2009; Linan-Thompson et al., 2006; Vaughn et al., 2006). The early identification and subsequent prevention of low reading comprehension and writing is the most effective method to reduce the prevalence of reading and writing difficulty in this population. It is quite evident that many young bilingual students need early, intensive oral language instruction that focuses on key oral academic language repertoires to foster grade-level reading comprehension and writing skills (Durlak, 1997; Gersten & Dimino, 2006).

To inform early, appropriate interventions for a large number of students (including bilinguals), an efficient, valid, and reliable screening and progress monitoring system is needed. Assessment tools that are commonly administered school-wide to young students tend to have a disproportionate focus on decoding (Silverman et al., 2019), and the validity of reading comprehension benchmark and progress monitoring instruments that are administered to young emerging readers is questionable (Moscovitch, 2004; Paleologos & Brabham, 2005; Petersen & Stoddard, 2018). Reading comprehension and writing cannot be measured in students who do not yet know how to read or write. What is needed is an alternative, proxy measure that validly assesses a similar construct to reading comprehension and writing in young pre-readers (Ukrainetz, 2006). Oral language, which can be assessed early, is an obvious candidate, especially if the oral language being assessed is as complex as the written language students are expected to read and understand. Without such measures, students cannot be identified as having reading comprehension and/or writing difficulty until problems have clearly emerged, making prevention impossible (Gersten & Dimino, 2006). Such an assessment would need to meet the stringent requirements of universal screening and progress monitoring measures (Deno, 2003; Deno et al., 1982). The preschool Narrative Language Measures-Listening (NLM-Listening) subtest of the CUBED assessment (Petersen & Spencer, 2016) is an oral language English and Spanish universal screening and progress monitoring tool that was designed to meet this purpose.

In this study, we examine the evidence of validity and reliability for the preschool NLM-Listening.

Language as a Tool for Assessing Comprehension and Production in Young Learners

There is considerable research evidence that supports using oral language as a proxy measure for reading comprehension and writing in young students. For example, it has been clearly established that successful reading comprehension relies heavily on a strong oral language foundation (Catts, et al., 2006; Elleman et al., 2009; Miller et al., 2006; Nation & Snowling, 1998; Nation et al., 2004; National Early Literacy Panel, 2008; Snow et al., 1998; Stothard & Hulme, 1992). Specifically, measures of oral vocabulary (Lervåg et al., 2010; National Reading Panel, 2000; Perfetti & Stafura, 2014) and narrative ability (Babayigit et al., 2021; Bishop & Edmondson, 1987; Fazio et al., 1996; Griffin et al., 2004; Mehta et al., 2005) are highly related to academic performance. Catts et al. (2006) obtained information on the language comprehension and decoding ability of 182 kindergarten children, then measured language comprehension, reading comprehension, and decoding ability when those same students were in second, fourth, and eighth grades. Results indicated that the children who were classified as having reading comprehension deficits in eighth grade also demonstrated language comprehension difficulties in kindergarten, second, and fourth grades. Their findings indicated that oral and written language comprehension were significantly intertwined, and students with language comprehension difficulties continued to have comprehension difficulties as grades increased, even when comprehension was measured differently across grades. Language comprehension weaknesses and strengths appear to persist across language modalities.

A Focus on Both Languages

Researchers recommend strengthening the foundation in the first language to better facilitate the development of a second language (Coltrane, 2003; Larson et al., 2020; MacSwan et al., 2017; Restrepo et al., 2013). For young Spanish-speaking students learning English, language instruction should be provided in both English and Spanish (Gutiérrez-Clellen, 1999; Peña & Kester, 2004; Restrepo et al., 2013; Thordardottir et al., 1997). Evidence suggests that instruction in a student's home language during the preschool years will lead to equivalent and, in some cases, superior English language proficiency and later academic achievement in comparison to English-only interventions (Kohnert et al., 2005; Larson et al., 2020; Restrepo et al., 2013). Universally screening all preschoolers' English language development through beginning, middle, and end of year assessments can help identify which children may benefit from language intervention, and adding a Spanish language assessment for children's whose home language is Spanish can lead to systematic support for their home language. This individualized attention to both languages can increase academic achievement and proficiency in the language of instruction and foster bilingualism.

Using Narratives to Assess Academic Language

Although oral language comprehension and production can be measured in a number of ways, the assessment of oral language that is most reflective of written language would most likely yield construct equivalence. The assessment of narrative language, which can be rife with complex academic language that reflects the written language which students are exposed in school, is a clear choice (National Governors Association Center for Best Practices and Council of Chief State School Officers, 2010; Petersen, 2011; Westby, 1985). Many have asserted that narratives bridge the gap between oral and written language (Gillam & Johnston, 1992; Spencer & Petersen, 2018; Westby 1985). The academic language in oral narratives is similar to the academic language in written narratives, which include specific, descriptive vocabulary, complex grammatical structures, and discourse elements such as character, setting, initiative event, attempt, consequence, resolution, (Hudson & Shapiro, 1991; Skarakis-Doyle & Dempsey, 2008). Narratives contain decontextualized language, which require a student to explain the context and referents to their audience (Dickinson & McCabe, 2001; Griffin et al., 2004). As has been evidenced in several studies, a young student's narrative language skills are highly predictive of their ability to understand and use complex, written language in later grades (Catts et al., 2002; Griffin et al., 2004). For example, Babayigit et al. (2021) found that five-year-old students' comprehension and narrative skills predicted reading comprehension at 10 and 14 years of age. It stands to reason that young children with good narrative skills experience fewer academic problems later (Bishop & Edmondson, 1987), and early storytelling abilities predict academic remediation in second grade (Fazio et al., 1996). A narrative retell language sample can also be particularly sensitive to receptive and expressive language changes over time because it captures considerable depth and breadth of a student's language ability. This sensitivity is important for detecting growth through frequent progress monitoring.

Multi-tiered System of Supports

To help young students with potential reading comprehension and writing difficulties, these students must first be accurately identified, appropriate interventions must be applied, and progress must be monitored. Multi-tiered systems of supports (MTSS) promise to fulfill such objectives through the use of universal screening (i.e., benchmarking), progress monitoring and tiered interventions. MTSS has the potential of greatly reducing assessment bias that plagues commonly used static measures, especially when administered to culturally and linguistically diverse students. This is because within MTSS, assessments are used dynamically, where the emphasis is on measuring what a student can learn when given instruction instead of measuring what a student currently knows (Petersen et al., 2017). This focus on response to intervention mitigates confounding variables, such as English language proficiency, prior education, and socioeconomic status (SES), which so very often render assessments uninterpretable. Furthermore, MTSS facilitates individualized instruction because the frequent monitoring of student progress does not go without action. When a student is not responding as expected with one tier of instruction, adjustments are made and more intensive instruction is provided. Flexibility and adaptability of interventions in an MTSS model require information gathered through repeated sampling of student behaviors. It is within the MTSS framework where repeated sampling of student performance is most needed. Yet there are few valid progress monitoring tools for oral language comprehension and production that can be used in an MTSS context for young students (Silverman et al., 2019; Wackerle-Hollman et al., 2021).

Deno, Mirking, and Chaing (1982) suggested that to be useful to judge response to intervention, assessment instruments need to have the following characteristics: a) measure authentic child behaviors and key skill elements that represent important outcomes, b) have multiple parallel forms so that observed changes are due to learning and development and not differences in instrumentation, c) have standardized administration and scoring procedures, d) be time efficient and economical, e) be sensitive to growth due to intervention or change over time, and f) meet the requirements of technical adequacy, including good reliability and validity evidence. The English and Spanish preschool NLM-Listening subtest of the CUBED assessment (CUBED; Petersen & Spencer, 2016) was carefully constructed to meet these characteristics (Deno, 2003; Fuchs, 2004; Petersen & Spencer, 2012). Importantly, it features the examiner reading a standard administration script, a unique story to the student, and asking the student to retell the story while the examiner scores using the story-specific rubric on the record form. In practice, each form takes approximately two minutes to administer and score. While there are NLM-Listening subtests for older grades (see Petersen et al., 2020), at this time, only the preschool level has a Spanish companion. Multiple parallel forms exist for each language so that a subset could be designated for universal screening at the beginning, middle, and end of the year and another subset could be designated for more frequent and regular probing to monitor students' comprehension and production of academic language over time and in relation to instruction and intervention.

To illustrate the intended use of the NLM-Listening, it is helpful to describe how it has been used in MTSS research. In an implementation study of a multitiered narrative-based academic language intervention in Head Start preschools (Spencer et al., 2017), the English version of the NLM-Listening was used to identify students who needed Tier 2 intervention and then to monitor their progress once intervention started. For the screening, children were administered three forms of the NLM-Listening in one session at each screening wave (beginning, middle, and end of year). Their retells were scored and the highest of the three scores were compared to the benchmark score of 8, which was established in previous research (Petersen & Spencer, 2016; Spencer et al., 2015). Students performing below that benchmark were assigned to receive Tier 2 small group intervention (scores of 2-7) or Tier 3 individual intervention (scores of 0-1) in addition to the Tier 1 instruction. In addition to delivering tiered instruction (once a week to the whole class) and intervention (twice a week to about 15% of students), once a month the Head Start teachers and teaching assistants administered one form of the NLM-Listening progress monitoring story set to the students receiving Tier 2/3 interventions. Their scores were plotted on a line graph and examined for growth over time and compared to the expected benchmark. Teachers used the scores and the information gleaned from listening to students' retells to make adjustments to intervention groups, frequency of sessions, and intervention targets. Spencer and colleagues also reported on the teachers' ability to administer the NLM-Listening with fidelity (mean = 96.6%) and to score it reliably (mean = 84.9%).

While promising evidence exists for earlier versions of the English NLM-Listening (Petersen & Spencer, 2012, 2016), there has been extensive recent development to include the Spanish NLM-Listening; therefore, additional research is needed to examine its potential to assess both English and Spanish language development of preschoolers and to be suitable for its intended purposes. In this study, we first evaluated evidence of alternate-forms reliability across the multiple forms in each language, supplemented by a factor analytic study of a subset of forms. We then examined evidence for external criterion validity by estimating the correlations between the Spanish or English retells and well-established criterion measures of oral language.

Method

Participants and Setting

A sample of 126 three- and four-year-old preschool students were drawn from five Head Start preschools in a Southwestern state. Researchers used parents' preferred language to explain the study procedures to parents when they dropped off or picked up their children from preschool. Parents signed a permission form and completed a brief demographic survey. Students were assigned to one of two samples—one sample for examining the reliability and validity evidence related to the English NLM-Listening (*N* = 65) and one sample for examining the reliability and validity evidence related to the Spanish NLM-Listening (*N* = 61)—informed by each parent's indication of home language on the survey and confirmation from the classroom teacher of the child's dominant language. Descriptions of the two participant groups are included in Table 1. Once the study began, all assessments and research activities took place in the Head Start preschool classrooms. While children participated in the NLM-Listening individually with an examiner, the other children in the classroom were engaged in learning centers.

Narrative Language Measures-Listening

The NLM-Listening for preschoolers involves an examiner reading a short story and asking the child to retell the story. To facilitate reading and scoring the story, each record form contains a standard administration script and instructions to the examiner, a unique story, and a story-specific scoring rubric (see Supplemental Material for an example). Each form of the NLM-Listening also has recall questions that can be administered following the examiner's reading of the story; however, we did not examine them in this study because they are optional and meant to supplement the retell information. Each version of the NLM-Listening (Spanish and English) has 25 forms, each based on a unique story. Nine of the 25 forms were randomly designated as benchmark assessments, whereas the remaining 16 forms were designated for more frequent administration to monitor children's language development. Stories feature child-relevant themes such as getting hurt, losing something, or wanting a toy someone else has. To facilitate equivalence and to ensure they contain comparable academic language, each story is approximately the same length (English = 68-70 words; Spanish = 73-75 words), has been leveled on story grammar, and contains the same number of uncommon words and complex clauses. For example, the stories have a main character who is named four times, a secondary character who is not named, a setting that includes an action and a location, a problem, a feeling, an attempt to solve the problem, a consequence, an ending, a temporal opening (e.g., the other day, last week), equivalent number and type of adjectives, adverbs, pronouns, conjunctions, and equivalent syntactic complexity with the same number of adverbial, nominal, and relative subordinate clauses. These academic language features reflect the complexity typical of written language, even though they were designed for oral language comprehension tasks. Importantly, Spanish stories were created based on Spanish language development, not English language development, and they are not translations of the English stories. Each story is accompanied by five simple illustrations representing main parts of the story.

Administration of the NLM-Listening was standardized. An examiner placed the picture book, showing five illustrations, in front of the child and said (in Spanish or English depending on the language of assessment), “I’m going to tell you a story. Please listen carefully. When I’m done, you are going to tell me the same story. Are you ready?” The examiner read the story word for word at a moderate pace with normal inflection. When the examiner was finished, they said, “Thanks for listening. Now you tell me that story.” Examiners recorded the students’ retells using a digital voice recorder. A single parallel form took approximately two minutes to administer.

When possible, examiners scored students’ retells as they were produced in real time. However, because young students often speak softly and the assessment was often administered under noisy conditions, the examiners listened to audio files at a later time to score when necessary. To score students’ retells, examiners used story-specific scoring rubrics that are included in the NLM-Listening record forms. Responses related to story grammar (e.g., character, setting, problem, feeling, attempt, consequence, and ending) that were clear and complete received two points, but incomplete or unclear responses received one point. Zero points were awarded if the element was missing from the student’s retell. When children produced stories with a clear and complete problem, attempt, consequence, or ending (episodic elements), additional points were awarded. The use of specific temporal and causal vocabulary words that suggest greater sentence complexity (i.e., *then, because, when,* and *after*) earned one point per use. *Because, when,* and *after* could earn up to three points, but the use of *then* could earn only one point maximum. A total retell score was calculated by summing points for story grammar, episode, and language complexity. If a child retold the story word for word and used three causal subordinate clauses and six temporal subordinate clauses, the total points possible would be 28; however, this maximum score is highly unlikely and not aligned with developmental expectations. For preschoolers, a score of eight indicates a minimally complete episode and a developmentally appropriate score (Spencer et al., 2015). Petersen and Spencer (2012) reported preliminary evidence of alternate form reliability for the English preschool NLM-Listening (.77, p <.001), with excellent fidelity of administration (91%). In addition, the story retell measure loaded on both comprehension and expression of language. Point-by-point inter-rater agreement for English preschool NLM-Listening stories that were scored in real-time was reported to be 91%, suggesting adequate interrater reliability.

External Criterion Measures

Two additional language measures were administered in each group's target language for the purpose of obtaining evidence of criterion-related validity for the NLM-Listening.

Clinical Evaluations of Language Fundamentals-Preschool. The Clinical Evaluations of Language Fundamentals-Preschool (CELF-P) in English (Wiig et al., 2004) and Spanish (Wiig et al., 2009) are norm-referenced tests designed to measure general language abilities. For this study, four subtests were administered: Word Structure, Sentence Structure, Expressive Vocabulary, and Following Directions. The reliability of these subtests in each language is adequate (English r = .77-.92; Spanish r = .69-.96). Administration lasted 20-25 minutes.

Natural language sample. Language samples collected in each group's target language (i.e., English or Spanish, respectively) through the use of a wordless picture book served as the second criterion measure. Students were shown a wordless picture book, *Frog Where Are You?* by Mercer Meyer, and listened to the examiner tell the story. When the examiner was finished, students retold the story. Students' retells were recorded and later transcribed. The transcriptions were entered into the Systematic Analysis of Language Transcripts (SALT) software (Miller & Iglesias, 2012), which analyzed each narrative for total number of words (TNW), number of different words (NDW), and mean length of utterance in words (MLU). These metrics are commonly used to characterize language productivity and complexity (Justice et al., 2006).

Research Design and Procedures

The procedures to administer the two NLM-Listening versions (English and Spanish) were identical, except for the language of assessment. Although from a design perspective, it would have been optimal to have each student complete all 25 forms in a language, as some did, we suspected some children would not be able to complete the entire set within the desired time frame (e.g., due to absences, fatigue, or distraction). To eliminate order effects and to randomize missing data on forms, a randomized Latin square design based on blocks of 25 was used to plan administration of the 25 forms in each participant group. In a full 25 student x 25 form Latin square, every test would have appeared once in each position, and each of the 25 NLM-Listening forms would have been administered to each student once. Additional Latin squares were generated and stacked to accommodate the maximum available sample size for each group. With this design, missing scores for children who were unable to receive or complete the full set of 25 forms were distributed randomly across the NLM stories and order effects due to practice were eliminated.

Based on the described design, unique sequences of the 25 forms were created for each participant and arranged in packets to ensure the order in which the assessments were administered was properly controlled. When testing began, students received 1-4 NLM-Listening forms a day, which depended on their cooperation (~2-12 minutes). When children expressed or displayed fatigue or disinterest (often after 2 or 3 forms), administration of the current form was terminated and the testing session would end. That form was not readministered at a later time and was considered missing. However, on the next day, the child would be administered the next form(s) in their packet. Regardless of how many forms each child had completed, the two criterion measures were administered after the child's 13th form in their packet of 25. Half the participants in each sample were randomly assigned to receive the CELF-P before a language sample was collected using the *Frog Where Are You?* wordless picture book, and the other half received the criterion measures in the opposite order. All assessments for an individual child, including up to 25 forms and both criterion measures, were administered within three to four weeks. However, if a child had not completed all 25 forms within four weeks, the research team did not attempt to administer the remaining forms in their packet.

Analytic Plan

Our analytic approach includes descriptive statistics and evidence to support an evaluation of the reliability and validity of the Spanish and English NLM-Listening. We first computed and examined descriptive statistics for the NLM-Listening retell scores and criterion measures, including an evaluation of the extent and potential impacts of missing scores. As the Spanish and English NLM-Listening versions are distinct sets of measures and are not direct translations, and, further, they were administered to different samples of students, we refrained from making direct statistical or descriptive comparisons between Spanish and English scores and focused on understanding the descriptive information for the forms within each language.

Next, we conducted analyses to examine whether the 25 NLM-Listening forms in each respective language can be used interchangeably. We computed alternate-forms reliability estimates based on the 300 pairwise correlations among the 25 forms in each language. We supplemented these alternate-forms reliabilities with a confirmatory factor analytic (CFA) study of a subset of the forms, specifically the 9 benchmark forms in each language designated for screening at the beginning, middle, and end of each academic year (3 forms at each time; Petersen & Spencer, 2016); the available sample size was insufficient to support simultaneous CFA evaluation of the full set of 25 forms. For the set of 9 benchmark forms in each language, we tested a series of three nested, progressively-constrained, unidimensional models that provided evidence regarding the degree of measurement equivalence across forms as follows (see Brown, 2015, pp. 207-221 for a detailed, accessible illustration of this approach): 1) congeneric forms that measure the same narrative language construct (i.e., all forms load on the same narrative language factor); 2) essentially tau-equivalent forms that have equivalent relationships with the narrative language construct such that a unit change in the latent narrative language construct is associated with the same amount of change in all alternate forms (i.e., equivalent factor loadings across forms); and 3) essentially parallel forms that additionally measure the narrative language construct with the same level of precision (i.e., equivalent factor loadings and error variances across forms). These nested models were compared with robust chi-square difference tests. Where necessary, modification indices were examined to aid in identifying parameters that were not equivalent across forms.

We then investigated whether the NLM-Listening assessed the intended narrative language construct. Specifically, we obtained external criterion validity by estimating the correlations between the Spanish or English retells and well-established criterion measures of oral language in each respective language.

Results

Descriptive Statistics for Spanish and English Retells

Descriptive statistics for Spanish and English retell scores are shown in Tables 2 and 3, respectively. Students completed 10-25 forms each of either Spanish or English NLM-Listening, with 91.6% of the 3,150 planned forms completed overall and high completion rates on individual forms. For the Spanish sample, completion rates for any one form ranged from 53 to 60 of the 61 total participants (i.e., see Valid *N* column in Table 2). For the English sample, completion rates for any one form ranged from 54 to 62 of the 65 total participants (i.e., see Valid *N* column in Table 3). No more than two participants shared any missing data pattern across the 25 forms. As a further probe of any systematic effects of missingness we also employed Missing Values Analysis in SPSS 28 to examine means and *t* tests ($\alpha = .05$) on the external criterion measures (CELF-P and FWAY) between those not missing vs. missing scores on each form; we summarize these findings broadly but do not present detailed results and caution against making inferences for any particular form given the means for the missing groups are estimated based on a very small number of cases (i.e., very few cases are missing for any one form). For all English and most Spanish form scores, missing status was unrelated to scores on the external criterion measures and almost all other forms; for five of the Spanish forms, missingness was slightly positively associated with students' performance on either the CELF-P or FWAY. Overall, students missing a score on a retell did not tend to have systematically higher or lower scores on the external measures or on other forms than those not missing the score. Given our use of the Latin square design to randomize administration of NLM-Listening forms, the relatively small amount of missing data, and evidence that missingness on a form was not consistently associated with student performance on external criterion measures or other NLM-Listening forms, we proceeded to compute descriptive statistics and pairwise correlations based on the observed data with what we expect to be minimal risk of bias.

As discussed in the Analytic Plan, we caution against making direct statistical or descriptive comparisons between Spanish and English NLM-Listening retell scores given they are distinct sets of measures and were administered to different samples. The distributions of all NLM-Listening forms had observed minimum scores of 0 and observed maximum scores ranging from 15-21 across the 25 forms in both languages.

Mean Spanish NLM-Listening scores ranged from 4.13 to 6.46 across the 25 forms (*SD*s = 4.67 to 6.24), and the median scores ranged from 2.50 to 7.00 (see Table 2). On most of the Spanish retells, approximately one-third of the sample had a score of 0, contributing to the positive skew of the score distributions for the Spanish NLMs. Across the 25 Spanish forms, the average skewness was .54, which may be regarded as moderately positive, but never exceeded 1 (range .32 - .96). In summary, the distributional characteristics of the 25 Spanish NLM scores were quite similar, which is reassuring, and we consider in the Discussion the relatively high proportion of zero scores.

Mean retell scores for the sample completing the English NLM-Listening ranged from 6.55 to 8.97 (*SD*s = 4.07 to 5.34), and the median scores ranged from 5.50 to 9.00 (see Table 3). On most of the English retells, approximately 10% of the sample had a score of 0, and skewness was minimal, averaging .20 (range -.25 to .51). The 25 distributions of English NLM-Listening scores were judged to be quite similar to each other, as we intended for these multiple-form assessments.

Evidence of Alternate-Form Reliability for NLM-Listening

In this section, we present evidence of alternate-form reliability for Spanish and English sets of NLM-Listening forms to inform whether the they may be regarded as interchangeable in practice. First, we computed all 300 pairwise correlations among the 25 retell scores in each language as alternate-form reliability estimates, which ideally would exceed approximately .70. Correlations were averaged by applying Fisher's *r* to *z′* transformation, computing the sample-size weighted mean of *z′*, and transforming the mean of *z′* back to *r*.

Second, CFAs were utilized to assess the degree to which the forms are parallel and may be used interchangeably. These CFAs were conducted on the subset of the 9 benchmark forms in each language that have been recommended for use in research and for screening in fall, winter, and spring of an academic year (3 forms at each time); the available sample size was insufficient to support simultaneous CFA evaluation of the full set of 25 forms. A robust maximum likelihood (MLR) estimator was used in Mplus 7.31 to handle missing data by using all available data from each participant to estimate model parameters and correct standard errors and fit statistics for non-normality. Nested models were compared with robust chi-square difference tests. The hypotheses implied by the unidimensional models are as follows: 1) congeneric forms measure the same narrative language construct (i.e., all forms load on the same factor, but loadings and residual variances are free to vary across NLMs; Model 1 for Spanish/Model 4 for English); 2) essentially tau-equivalent forms have equivalent relationships with the narrative language construct such than a one-unit change in the narrative language construct yields the same amount of change in all benchmark forms (i.e., factor loadings are constrained to be equivalent across forms except where noted below and in Table 4; Models 2a and 2b for Spanish/Model 5 for English); and 3) essentially parallel forms additionally constrains the narrative language construct to be measured with the same level of precision (i.e., factor loadings and residual variances are constrained to be equivalent across forms except where noted below and in Table 4; Models 3a-3c for Spanish; Model 6 for English).

Reliability Evidence for Spanish Retells. All 300 pairwise correlations among the 25 Spanish retell scores were computed as alternate-form reliability estimates, with pairwise sample sizes for Spanish retell scores ranging from 47 to 60. A histogram of these 300 alternate-form correlations for Spanish retells are displayed on the left side of Figure 1. Both the corrected mean (computed using Fisher's *r* to *z′* transformation) and the median of the Spanish alternate-form reliability estimates were .80 and are considered strong, with means ranging from .58 to .90. Only 10/300 (3%) reliability estimates were less than .70; four of these lowest 10 correlations involved Form 1.

Next, we tested the described set of CFAs to assess the degree to which the 9 benchmark Spanish forms are congeneric, tau equivalent, or parallel as evidence of measurement across forms. The congeneric model (i.e., one-factor, loadings free to vary across forms), Model 1, fit the Spanish Form 1-9 data well, and all forms loaded significantly on the narrative language factor. The essentially tau-equivalent model, Model 2a (i.e., set-wise test of equivalent factor loadings), was also tenable based on favorable model fit indices (except a larger SRMR) and the non-significant chi-square difference test comparing Model 2a (i.e., all loadings constrained to be equal) with Model 1 (i.e., all loadings freely estimated), supporting that the set of 9 Spanish NLM-Listening forms had equivalent relationships with the narrative language construct. We observed, however, that the SRMR and the upper bound of the 90% CI for the RMSEA both increased, indicating that constraining all loadings to be equal induced some lack of fit. A large modification index for the loading of Form 1, and the recognition that power for detecting differences in model fit may be low due to small sample size led us to decide this loading should not be regarded as equivalent; accordingly, the loading for Form 1 was allowed to vary from those of the other 8 loadings in Model 2b, which demonstrated good fit and a statistically significant improvement over model 2a. In the next step, Model 3a was specified to reflect essentially parallel forms by applying set-wise constraints on both loadings (as in Model 2a) and residual variances to test the hypothesis that narrative language was measured with the same level of precision on all 9 forms. The fit of the fully parallel form Model 3a, although adequate according to the RMSEA and model chi-square, showed a statistically significant decline in comparison with tau-equivalent Model 2a. Sequential consideration of modification indices suggested that some degree of non-equivalence of residual variances was detected on Form 1 (Model 3b freed residual variance for Form 1) and, in the next step and to a lesser extent, Form 2 (Model 3c freed residual variances for Forms 1 and 2), indicating the measurement precision of these two forms may vary from that of the other Spanish benchmark forms. Parameter estimates for the final Spanish NLM-Listening model (Model 3c) are shown in the top panel of Figure 2; all standardized factor loadings were strong (≥.82) and statistically significant.

In summary, alternate-form reliability estimates were consistently strong across the 25 Spanish NLM forms, averaging .80. Further, a rigorous psychometric analysis of the 9 Spanish benchmark forms showed that these may forms be regarded as: a) measuring the same narrative language construct (congeneric measures, Model 1); b) having equally strong relationships with the narrative language construct such than a one-unit change in the narrative language construct yields the same amount of change in all benchmark forms, with the exception of Form 1 (tau-equivalent measures, Model 2b); and 3) measuring narrative language additionally with the same level of precision across forms, again with the exception of Form 1 and possibly Form 2 (parallel forms, Models 3b and 3c, respectively).

Reliability Evidence for English Retells. All 300 pairwise correlations were computed on English retells as well; pairwise sample sizes for English retell scores ranged from 47 to 62. The right side of Figure 1 displays a histogram of these 300 alternate-forms correlations for English retells. The median of the alternate-forms reliability estimates for English retells was .63 and the mean was .64 (computed using Fisher's *r* to *z′* transformation), with means ranging .46 to .82. Overall, 17.3% of the 300 correlations exceeded .70 and 68.3% were greater than .60. Only 4 correlations (1.3%) were less than .50; two of these involved Form 2.

The results of comparisons among progressively constrained CFA models for the 9 benchmark English retells are shown in Table 4. The congeneric (i.e., one-factor) model marginally fit the English Form 1-9 data (Model 4), and all forms loaded significantly on the narrative language construct. Regarding fit, the nonsignificant model chi-square test of perfect fit and the SRMR value were favorable, and the CFI and RMSEA were at the margin of acceptable fit, with the 90% CI likely being wider than desired due to the small sample size. The test of Model 5, with equated factor loadings reflecting the essential tau-equivalence hypothesis, held (i.e., non-significant chi-square difference test for Models 5 vs. 4), supporting that the set of 9 English NLM-Listening forms had equivalent relationships with the narrative language construct; model fit remained similarly marginal. Finally, the test for essentially parallel forms in Model 6, with equality constraints across forms on loadings and residual variances, also held (i.e., non-significant chi-square difference test of Models 6 vs. 5), supporting equal measurement precision across the 9 forms. Parameter estimates for the final English model for parallel forms

(Model 6) are shown in the bottom panel of Figure 2; standardized factor loadings were strong (.80) and statistically significant.

In summary, the alternate-form reliability estimates were somewhat variable across the 25 English NLM forms and tended to be a bit lower than desired, averaging .64. Additional psychometric analysis via progressively-constrained, unidimensional CFA models focusing on the 9 English benchmark forms showed that these may forms be regarded as: a) measuring the same narrative language construct (congeneric measures, Model 4); b) having equivalent relationships with the narrative language construct such than a one-unit change in the narrative language construct yields the same amount of change in all benchmark forms (tau-equivalent measures, Model 5); and 3) measuring narrative language additionally with equivalent precision across forms (Models 6).

Evidence of Validity for NLM-Listening

Validity Evidence for Spanish Retells. The summary statistics for correlations of scores on the 25 forms of Spanish retells with CELF-P and FWAY language measures are shown in Table 5. Correlations between the NLM-Listening retells and the external criterion scores were averaged appropriately across the 25 forms by applying Fisher's *r* to *z*′ transformation, computing the sample size weighted mean of *z*′, and transforming the mean of *z*′ back to *r*. Additionally, form-specific correlations with the CELF-P total score and the FWAY total word score are presented in last two columns of Table 2, and form-specific correlations with the other CELF-P and FWAY scale scores are available from the first author. All 25 Spanish forms were positively correlated with each of the CELF-P and FWAY scores. Average correlations of Spanish retells with the external measures were moderate to strong in magnitude, ranging on average from .46 for the FWAY mean length of utterance (FWAY-MLU) to .74 for the CELF sentence structure scores (CELF-SS). Correlations of Spanish retells with FWAY measures evidenced somewhat greater variability and ranges across the 25 forms than the CELF-P scores.

Given that 25 correlations were examined for each external validity criterion, an alpha of .05/25 = .002 was applied to evaluate statistical significance of each correlation. After applying the corrections to significance levels, all 25 Spanish forms with each of the five CELF-P measures, the FWAY-total, and the FWAY-NDW measure were statistically significant at the .002 level (and all but Form 10 was also significant at the .001 level). For FWAY-NTW, correlations with all Spanish forms were statistically significant at the .002 level except for Form 18, $r = .41$, $p = .003$. For FWAY-MLU, correlations with 17 Spanish retells were statistically significant at the .002 level, whereas 8 were not. Spanish Forms 10 and 18 showed somewhat weaker evidence of criterion-related validity than the other 23 forms based on their lower correlations with more than one of the nine external measures. In summary, across the nine criterion measures, these findings offer strong evidence of criterion-related validity for the Spanish NLM-Listening.

Validity Evidence for English Retells. Summary statistics for correlations of scores on the 25 forms of English retells with CELF-P and FWAY language measures are displayed in Table 6, in which average correlations were computed as described for the Spanish retells. English form-specific correlations with the CELF-P total score and the FWAY total word score are presented in the last two columns of Table 2, and form-specific correlations with the other CELF-P and FWAY scale scores are available from the first author. Estimated correlations of all 25 English forms with each of the CELF-P and FWAY scores were positive. Correlations of English retells with these external language measures were moderate in magnitude, ranging on average between .41 and .53, with the exception of FWAY mean length of utterance (FWAY-

MLU) and total (FWAY-total), which on average correlated lower with NLM-Listening retells at .28 and .29, respectively.

As the standard deviations and ranges of correlations in Table 6 show, there was some variability in the strength of external-criterion validity evidence across the 25 English NLM-Listening forms. Applying an adjusted Type I error rate of α=.002, correlations of all 25 English forms with CELF-FD were statistically significant (also significant at the .001 level). For both CELF-Total and CELF-SS, correlations with all but three English forms were statistically significant. Nine and seven forms of the English NLM-Listening were not significantly correlated at the α=.002 level with CELF-WS and CELF-EV, respectively. Across the set of correlations of the 25 English forms with the five CELF-P scores, only four English forms had nonsignificant correlations at the α=.002 level with two or more of the external CELF-P measures (number of nonsignificant correlations): Form 1 (4); Form 12 (3); Form 15 (2); and Form 23 (4).

With respect to correlations with FWAY scores, FWAY-NDW and FWAY-NTW had statistically significant correlations at the α=.002 level with 19 and 15 English forms, respectively. FWAY-MLU and FWAY-total, for which the estimated correlations with English forms were much lower, had only three and four forms, respectively, that evidenced statistically significant correlations with these measures. Across the set of correlations with the four FWAY scores, 18 English forms had nonsignificant correlations at the α=.002 level with two or more external FWAY measures. In summary, English NLM-Listening Forms 1, 12, 15, and 23 showed somewhat less evidence for validity than the others based on having two or more nonsignificant correlations with measures in both the CELF and FWAY sets of external criterion measures. For the English language sample, correlations of the NLM-Listening forms with seven of the nine criterion measures offered moderately strong evidence of criterion-related validity; relationships with FWAY-total and FWAY-MLU were somewhat weaker.

Discussion

Given the overwhelmingly high percentage of bilingual students who struggle with English reading comprehension across the U.S. (NAEP, 2019), there is a clear need to develop an efficient, valid, and reliable proxy measure that educators can use to monitor the development of academic oral language in pre-readers (Kieffer & Vukovic, 2012; Silverman et al., 2019). This assessment would need to meet the requirements of general outcome measurement to be used in MTSS (Deno et al., 1982). The purpose of this study was to examine evidence of reliability and validity of the preschool English and Spanish NLM-Listening subtests of the CUBED assessment (Petersen & Spencer, 2016). Specifically, we examined alternate-form reliability using pairwise correlations and confirmatory factor analysis and investigated criterion-related concurrent evidence of validity and construct validity through the examination of unidimensionality across forms via exploratory factor analysis. Our results indicated that the mean alternate-form reliability estimates were .80 for the Spanish NLM-Listening and .64 for the English NLM-Listening. Within both English and Spanish language sets, the 9 forms used for benchmark universal screening had strong standardized factor loadings and evidenced reasonably equivalent relationships with the narrative language factor, and they measured narrative language with similar precision with only a few exceptions. Evidence of validity was also obtained through concurrent validity estimates, which suggested that the NLM-Listening measured similar oral language constructs as the CELF-P and the FWAY narrative elicitation procedures.

Reliability Evidence across Alternate Forms

The purpose of alternate forms is to mitigate the testing effect encountered when the same test is administered multiple times. Differences in scores from one administration to the next should be primarily attributable to the student's change in skill or knowledge, not due to factors such as one form being easier than the other. Because language performance is inextricably tied to culture, dialect, and vocabulary exposure, it is inevitable that individual stories will resonate with some students more than others. In addition, the administration and scoring of the NLM-Listening are not perfectly reliable. The variations in story topic, vocabulary, administration, and scoring, even when all other features are held constant, means that perfect alternate form reliability is impossible (Cohen & Swerdlik, 2018).

The differences observed between the English and Spanish NLM-Listening in the current study could be reflective of the language differences in the populations. The English NLM-Listening was examined with a monolingual English-speaking group of preschoolers and the Spanish NLM-Listening was examined with Spanish-English sequential bilingual children. It is also possible that the reliability estimates were higher among the Spanish forms because of floor effects. As these children were young and learning two languages (most frequently one at home and one at school), it is possible that the NLM-Listening was not sensitive enough to detect the bilingual children's emerging Spanish skills.

The question is whether alternate forms with construct equivalency and mean correlations among forms of .80 and .64 for Spanish and English assessments, respectively, are adequate for screening and progress monitoring within MTSS, the purposes for which the NLM-Listening was designed. The answer to this depends on the extent to which performance on the NLM-Listening is expected to improve from one time point to the next. When it is administered to preschoolers in conjunction with narrative-based language intervention, moderate to large effect sizes are typically reported from pre- to post-intervention (Spencer et al., 2015; Spencer et al., 2019; Spencer et al., 2020). This means that even if the NLM-Listening forms are not perfectly parallel, differences in performance between the forms over time is likely attributable to an increase in student skill or knowledge. Furthermore, a consistent pattern of growth with multiple data points can provide convincing evidence that the progress is a within-child factor (Fuchs & Fuchs, 2008). This type of validity is available from several efficacy studies, which show that the NLM-Listening is sensitive to change over time.

Although we consider the alternate forms to be adequate, to improve alternate form reliability, the first logical step would be to examine any outlying forms and adjust those forms accordingly. In the case of the preschool NLM-Listening, Form 1 in Spanish and Form 2 in English may require revisions. This was corroborated by the results of the confirmatory factor analysis which indicated that although a one factor, congeneric model fit the data well for the 9 benchmark measures in each language, some degree of non-equivalence of factor loadings and residual variances was detected on Form 1 in Spanish. A post-hoc examination of Form 1 revealed that the main and secondary characters are introduced in the same beginning sentence, which is a feature not found in other forms (or stories). This modified structure impacted other parts of the Form 1 story as well. Therefore, revisions to this specific story will likely improve equivalency.

Concurrent, Criterion-Related Evidence of Validity

Concurrent, criterion-related evidence of validity was moderate to strong for the Spanish and English NLM-Listening. We used the CELF-P, a commonly administered norm-referenced test of language, and the FWAY narrative retell, a commonly used language sample elicitation procedure, as criterion measures. We did not consider either the CELF-P or the FWAY to be absolute gold standard assessment tools for English/Spanish dual language learners (Denman at al., 2017). Nevertheless, we would expect at least moderate correlations between the CELF-P or the FWAY with the NLM-Listening to demonstrate that the NLM-Listening measures related constructs.

The CELF-P and the NLM-Listening differ at a fundamental level. They are constructed based on different concepts of the dimensionality of language and the tasks used to measure those dimensions differ greatly. The CELF-P is comprised of subtests designed to measure different aspects of expressive or receptive language using tasks mostly separated from conversation or discourse. Although traditionally expressive and receptive language have been considered to be different modalities, recent research has brought to light that such a distinction is not accurate, and language in this age group is more unidimensional (LARRC, 2015; Tomblin & Zhang, 2006). Furthermore, when language is fractionalized, a student's ability to integrate all aspects of language in a meaningful communicative context is not assessed. The preschool NLM-Listening, on the other hand, is a discourse task that requires children to listen to and retell a model narrative through the integration of story grammar, vocabulary, and syntax. Thus, the retell scores from the NLM-Listening would most likely reflect different aspects of language when compared to the CELF-P, yet moderate to strong correlations should still exist, which is in alignment with the results of this study.

The FWAY and the NLM-Listening language elicitation tasks are quite similar, although the FWAY task uses a wordless picture book and a considerably longer model narrative. Despite these differences, we expected correlations between the FWAY and the NLM-Listening to be strong. Although the results of this study indicated that there were moderate to strong correlations between the two narrative language elicitation tasks, it is possible that correlations between the FWAY and the NLM-Listening would have been stronger had we scored the story grammar from the FWAY retells. Because the FWAY retells were not analyzed for story grammar complexity, the data from those retells were potentially truncated, resulting in less variance and weaker correlations. It is also possible that correlations would have been stronger had syntax and vocabulary been scored in the same manner. Recall that the NLM-Listening scores were derived from story grammar, episode complexity (a sub-analysis of story grammar), and the frequency of use of subordinating conjunctions (reflecting both vocabulary and syntax). The FWAY analyses were conducted using the SALT software, which provided vocabulary information through TNW and NDW and syntax information through MLU and a subordination index. Furthermore, retells elicited using the FWAY were derived from a model narrative that lacked the same degree of condensed complex academic language as the NLM-Listening stories. Hence, our findings that correlations between the NLM-Listening and the FWAY were in the moderately-strong range were not surprising. Nonetheless, this evidence of validity indicates that the NLM-Listening measures similar constructs as those reflected in the FWAY results.

Spanish NLM-Listening Forms 10 and 18 showed somewhat weaker evidence of criterion-related validity than the other 23 forms based on their lower correlations with more than one of the external measures. These forms need to be examined further to determine why they yield somewhat outlying results. For the English language assessments, correlations of the forms with seven of the nine criterion measures offered moderately strong evidence of criterion-related validity for the English NLM-Listening, with relationships with FWAY-total and FWAY-MLU being somewhat weaker. However, for both English and Spanish, concurrent, criterion-related evidence of validity supported the accumulating evidence of the validity of the NLM-Listening as an oral language measure appropriate for progress monitoring.

Application within MTSS

Because the NLM-Listening was designed for assessment within MTSS contexts, which is most likely to occur in English in the U.S., the English NLM-Listening would be the best candidate for identifying children with English language intervention needs (i.e., universal screening). There are no assessments that have perfectly equivalent alternate forms. The NLM-Listening is no exception. However, there is a standard recommendation for accounting for the variance between forms in practice (Barth et al., 2012). When used for universal screening, early educators can administer three forms to preschoolers in a single session and use the highest score of the three for determining the child's need for Tier 2 intervention. This strategy increases the stability and validity of that score without necessitating perfectly equivalent forms. In MTSS frameworks, universal screening typically takes place at the beginning, middle, and end of the year, which would require nine NLM-Listening forms. For children whose home language is Spanish, the Spanish NLM-Listening could serve as a supplemental source of data collected at those universal screening timepoints to assist in intervention planning, indicating whether a monolingual or bilingual intervention would be most appropriate and revealing a child's relative language strengths and weaknesses. The English and Spanish NLM-Listening has been used in previous MTSS research in exactly this manner (Spencer et al., 2019; Spencer et al., 2020). It should be noted that a benchmark score has already been established for the English version (Petersen et al., 2014; Petersen & Spencer, 2012). To be able to use the Spanish version by itself for identification, however, additional research is needed to establish cut scores. This may necessitate a significant amount of work because U.S. Spanish-speaking children comprise an incredibly heterogenous group and cut scores for each subgroup would likely be needed.

The second purpose of designing the NLM-Listening assessment with 16 additional parallel forms is to allow for repeated assessment of language over time vis-à-vis intervention. If monolingual English-speaking children receive an English only intervention, early educators should administer one English NLM-Listening every week, two weeks, or month to evaluate the extent to which the intervention is working. Trends in retell scores over time offer critical information about how potent the intervention is, what intervention modifications might be helpful, and when more (Tier 3) or less (Tier 1) intensive instruction for a child is warranted. For that reason, only one form needs to be administered at regular time points. If children receive intervention exclusively in Spanish or in both English and Spanish, the Spanish NLM-Listening could be used to monitor their Spanish language development and to inform intervention planning, modification, and potency (see Spencer et al., 2019 for an example).

The more stringent criteria, including multiple parallel forms, are needed for MTSS implementation; however, there are other contexts in which the NLM-Listening can be used that do not necessitate the full set of 25 forms. In pretest-posttest group research, the NLM-Listening in English and/or Spanish can assist in the detection of intervention efficacy (as in Spencer et al., 2020). If used in this manner, the most valid representation of children's true language abilities is needed. Therefore, three forms should be administered at each timepoint to account for the inevitable variance. In practice, speech-language pathologists and special education teachers can use the NLM-Listening to track their clients' progress with respect to targeted goals or to inform the differentiation in small group interventions with a variety of clients. For quarterly progress reports, administration of three might be needed, but more frequent (weekly, biweekly, or monthly) assessment would only require one NLM-Listening at a time.

Limitations and Future Directions

Collecting data from students on the full set of 25 forms in each language was an intensive undertaking and, accordingly, the available sample sizes in each group (i.e., 61 for Spanish retells and 65 for English retells, respectively) were somewhat smaller than might be desired for conducting particular psychometric analyses on the full set of measures simultaneously. For example, the available sample sizes were insufficient to support evaluation of the progressively constrained congeneric, essentially tau-equivalent, and essentially parallel models on the full set of 25 forms, so these evaluations were conducted on the 9 versions designated as benchmark forms in each language. We had sufficient power to detect two non-parallel forms (i.e., Spanish Forms 1 and 2), but it is possible that with larger samples and greater power, some degree of non-equivalence may be detected for other forms. Further research with a larger sample may support a more complete evaluation of these findings, as well as supporting the assessment of any future revisions to forms recommended by findings in the present study.

Further validation is needed with children from different SES backgrounds and demographics. Note that this sample was all from Head Start and the dual language learners were mostly from Mexican American backgrounds. Different SES, education levels, and Latin-American backgrounds would be more representative and generalizable. In addition, this study did not examine whether the measures are sensitive to change over time. Although we have partial validity data for this purpose with the intervention studies, further validation is needed.

We also recommend that future research more closely inspects the differences in reliability and validity results between the English and Spanish versions of the NLM-Listening. As they are distinct assessments and we examined each of them with distinct groups of children, close alignment was not expected. Substantial floor effects for the Spanish NLM-Listening were most likely the reason for the observed differences in the current study. The restricted range potentially increased the reliability coefficients obtained. However, this is only speculation at this point. This aspect should be investigated with more precision, potentially administering both versions to the same group of students.

Conclusions

The results of this study support the use of the preschool English and Spanish NLM-Listening assessment as a screening and progress monitoring tool that may serve as a proxy measure for the comprehension of written academic language. Although the evidence is not perfect nor as high as desired, it is important to keep in mind that the construct measured in the NLM-Listening is the oral language. It would be inappropriate to expect reliability for higher order oral language skills such as story retelling to be as high as simple, discrete, and easily countable skills such as picture naming or words read correctly. Given this consideration, evidence for alternate forms largely supports interchangeable use of NLM-Listening forms, albeit with minor revisions to some of the stories and the use of more than one form to make practical identification and intervention decisions. Having valid and reliable parallel forms for language is critical for early intervention planning, especially for young children who do not yet read, as oral narrative skill predicts reading comprehension and academic achievement (Catts et al., 2002; Dickinson & McCabe, 2001). Such valid and reliable forms, which are also easy and efficient to administer and score, easily fit within MTSS.

References

*Authors ’publications have been removed for blinding.

Babayigit, S., Roulstone, S., & Wren, Y. (2021) Linguistic comprehension and narrative skills predict reading ability: A 9-year longitudinal study Selma. *British Journal of Educational Psychology*, *94*, 148–168. https://doi.org/10.1111/bjep.12353

Barth, A. E., Stuebing, K. K., Fletcher, J. M., Cirino, P. T., Romain, M., Francis, D., & Vaughn, S. (2012). Reliability and validity of oral reading fluency median and mean scores among middle grade readers when using equated texts. *Reading psychology*, *33*(1-2), 133-161. https://doi.org/10.1080/02702711.2012.631863

Bishop, D. V., & Edmondson, A. (1987). Language-impaired 4-year-olds: Distinguishing transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52*(2), 156-173. https://doi.org/10.1044/jshd.5202.156

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). New York, NY: The Guilford Press.

Catts, H. W., Adolf, S. M., & Weismer, S. E. (2006). Language deficits in poor comprehenders: A case for the simple view of reading. *Journal of Speech, Language, and Hearing Research, 49*(2), 278-293. https://doi.org/10.1044/1092-4388(2006/023)

Catts, H. W., Fey, M. E., Tomblin, J. B., & Zhang, X. (2002). A longitudinal investigation of reading outcomes in children with language impairments. *Journal of Speech, Language, and Hearing Research, 45*(6), 1142-1157. https://doi.org/10.1044/1092-4388(2002/093)

Cirino, P. T., Vaughn, S., Linan-Thompson, S., Cardenas-Hagan, E., Fletcher, J. M., & Francis, D. J. (2009). One-year follow-up outcomes of Spanish and English interventions for

English language learners at risk for reading problems. *American Educational Research Journal, 46*(3), 744-781. https://doi.org/10.3102/0002831208330214

Cohen, R. J. & Swerdlik, M. E. (2018). *Psychological testing and assessment: An introduction to tests and measurement* (9th ed.). New York, NY: McGraw-Hill.

Coltrane, B. (2003). *Working with young English language learners: Some considerations.* ERIC Digest. https://files.eric.ed.gov/fulltext/ED481690.pdf

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y. W., & Cordier, R. (2017). Psychometric properties of language assessments for children aged 4–12 years: A systematic review. *Frontiers in Psychology*, *8*, 1515. https://doi.org/10.3389/fpsyg.2017.01515

Deno, S. L. (2003). Developments in curriculum-based measurement. *The Journal of Special Education, 37*(3), 184-192. https://doi.org/10.1177/00224669030370030801

Deno, S. L., Mirkin, P. K., & Chiang, B. (1982). Identifying valid measures of reading. *Exceptional Children, 49*, 36-45. https://doi.org/10.1177/001440298204900105

Dickinson, D. K., & McCabe, A. (2001). Bringing it all together: The multiple origins, skills, and environmental supports of early literacy. *Learning Disabilities Research & Practice*, 16(4), 186-202. https://doi.org/10.1111/0938-8982.00019

Durlak, J.A. (1997). Primary prevention programs in schools. In: Ollendick T.H. & Prinz R.J. (Eds.), *Advances in Clinical Child Psychology: Vol 19.* Springer, Boston, MA. https://doi.org/10.1007/978-1-4757-9035-1_8

Elleman, A. M., Lindo, E. J., Morphy, P., & Compton, D. L. (2009). The impact of vocabulary instruction on passage-level comprehension of school-age children: A metaanalysis. *Journal of Research on Educational Effectiveness, 2*(1), 1-44. https://doi.org/10.1080/19345740802539200

Fazio, B. B., Naremore, R. C., & Connell, P. J. (1996). Tracking children from poverty at risk for specific language impairment: A 3-year longitudinal study. *Journal of Speech, Language, and Hearing Research, 39*(3), 611-624. https://doi.org/10.1044/jshr.3903.611

Fuchs, L. S., & Fuchs, D. (2008). The role of assessment within the RTI framework. In D. Fuchs, L. S. Fuchs, & S. Vaughn (Eds.), *Response to intervention: A framework for reading educators* (pp. 27–49). Newark, DE: International Reading Association.

Fuchs, L. S. (2004). The past, present, and future of curriculum-based measurement research. *School Psychology Review*, *33*(2), 188-192. https://doi.org/10.1080/02796015.2004.12086241

Gersten, R., & Dimino, J. A. (2006). RTI (Response to Intervention): Rethinking special education for students with reading difficulties (yet again). *Reading Research Quarterly*, *41*(1), 99-108. https://doi.org/10.1598/rrq.41.1.5

Gillam, R. B., & Johnston, J. R. (1992). Spoken and written language relationships in language/learning-impaired and normally achieving school-age children. *Journal of Speech, Language, and Hearing Research, 35*(6), 1303-1315. https://doi.org/10.1044/jshr.3506.1303

Greenhalgh, K. S., & Strong, C. J. (2001). Literate language features in spoken narratives of children with typical language and children with language impairments. *Language, Speech, and Hearing Services in Schools, 32*(2), 114-125. https://doi.org/10.1044/0161-1461(2001/010)

Griffin, T. M., Hemphill, L., Camp, L., & Wolf, D. P. (2004). Oral discourse in the preschool years and later literacy skills. *First Language*, *24*(2), 123-147. https://doi.org/10.1177/0142723704042369

Gutiérrez-Clellen, V. F. (1999). Language choice in intervention with bilingual children. *American Journal of Speech-Language Pathology*, *8*(4), 291-302. https://doi.org/10.1044/1058-0360.0804.291

Hudson, J. A., & Shapiro, L. R. (1991). From knowing to telling: The development of children's scripts, stories, and personal narratives. In A. McCabe & C. Peterson (Eds.), *Developing Narrative Structure* (pp. 89-136). Hillsdale, NJ: Lawrence Erlbaum.

Justice, L. M., Bowles, R. P., Kaderavek, J. N., Ukrainetz, T. A., Eisenberg, S. L., & Gillam, R. B. (2006). The index of narrative microstructure: A clinical tool for analyzing school-age children's narrative performances. *American Journal of Speech-Language Pathology, 15*, 177-191.

Kieffer, M. J., & Vukovic, R. K. (2012). Components and context: Exploring sources of reading difficulties for language minority learners and native English speakers in urban schools. *Journal of Learning Disabilities, 45*(5), 433-452. https://doi.org/10.1177/0022219411432683

Kohnert, K., Yim, D., Nett, K., Kan, P. F., & Duran, L. (2005). Intervention with linguistically diverse preschool children. *Language, Speech & Hearing Services in Schools*, *36*(3), 251-263. https://doi.org/10.1044/0161-1461(2005/025)

Labov, W. (1972). *Sociolinguistic patterns* (No. 4). University of Pennsylvania press. https://doi.org/10.1017/s0047404500004528

Language and Reading Research Consortium (LARRC). (2015). The dimensionality of Spanish in young Spanish-English dual-language learners. *Journal of Speech, Language, and Hearing Research, 58*(3), 754-766. https://doi.org/10.1044/2015_JSLHR-L-13-0266

Larson, A.L., Cycyk, L., Carta, J., Hammer, C.S., Baralt, M., Uchikoshi, Y., Gigi An, Z, & Wood, C. (2020). A systematic review of language-focused interventions for young children from culturally and linguistically diverse backgrounds. *Early Childhood Research Quarterly, 50*(1), 157-178. https://doi.org/10.1016/j.ecresq.2019.06.001.

Lervåg, A. & Grøver Aukrust, V. (2010). Vocabulary knowledge is a critical determinant of the difference in reading comprehension growth between first and second language learners. *The Journal of Child Psychology and Psychiatry, 51*(5), 612-620. https://doi.org/10.1111/j.1469-7610.2009.02185.x

Linan-Thompson, S., Vaughn, S., Prater, K., & Cirino, P. T. (2006). The response to intervention of English language learners at risk for reading problems. *Journal of Learning Disabilities, 39*(5), 390-398. https://doi.org/10.1177/00222194060390050201

MacSwan, J., Thompson, M. S., Rolstad, K., McAlister, K., & Lobo, G. (2017). Three theories of the effects of language education programs: An empirical evaluation of bilingual and English-only policies. *Annual Review of Applied Linguistics, 37*, 218-240. https://doi.org/10.1017/s0267190517000137

Mehta, P. D., Foorman, B. R., Branum-Martin, L., & Taylor, W. P. (2005). Literacy as a unidimensional multilevel construct: Validation, sources of influence, and implications in a longitudinal study in grades 1 to 4. *Scientific Studies of Reading*, *9*(2), 85-116. https://doi.org/10.1207/s1532799xssr0902_1

Miller, J. & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research

Version 2012 [Computer Software]. https://www.saltsoftware.com/

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. https://doi.org/10.1111/j.1540-5826.2006.00205.x

Moscovitch, E. (2004). *Evaluation of the Alabama Reading Initiative*. Montgomery, AL: Cape Ann Economics for the Alabama State Department of Education. https://eric.ed.gov/?id=ED464366

Nakamoto, J., Lindsey, K. A., & Manis, F. R. (2007). A longitudinal analysis of English language learners' word decoding and reading comprehension. *Reading and Writing, 20*(7), 691-719. https://doi.org/10.1007/s11145-006-9045-7

Nation, K., & Snowling, M. J. (1998). Semantic processing and the development of word-recognition skills: Evidence from children with reading comprehension difficulties. *Journal of Memory and Language, 39*(1), 85-101. https://doi.org/10.1006/jmla.1998.2564

Nation, K., Clarke, P., Marshall, C. M., & Durand, M. (2004). Hidden language impairments in children. *Journal of Speech, Language, and Hearing Research, 47*(1), 199-211. https://doi.org/10.1044/1092-4388(2004/017)

National Early Literacy Panel (US). (2008). *Developing early literacy: Report of the National Early Literacy Panel: A scientific synthesis of early literacy development and implications for intervention*. National Institute for Literacy. https://doi.org/10.1037/e563852009-001

National Governors Association for Best Practices, Council of Chief State School Officers. (2010). *Common Core State Standards*. Washington D.C.

National Reading Panel (US), National Institute of Child Health, & Human Development (US). (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups*. National Institute of Child Health and Human Development, National Institutes of Health.

Paleologos, T. M., & Brabham, E. G. (2005). The effectiveness of DIBELS oral reading fluency as a predictor of reading comprehension for high-and low-income students. *Reading Psychology, 32*(1), 54-74. https://doi.org/10.1080/02702710903341262

Peña, E. D., & Kester, E. S. (2004). Semantic development in Spanish-English bilinguals: Theory, assessment, and intervention. In B. A. Goldstein (Ed.), *Bilingual language development and disorders in Spanish-English speakers* (pp. 105–128). Paul H Brookes Publishing.

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. *Scientific Studies of Reading*, *18*(1), 22-37. https://doi.org/10.1080/10888438.2013.827687

Petersen, D. B. (2011). A systematic review of narrative-based language intervention with children who have language impairment. *Communication Disorders Quarterly, 32* (4), 207–220. 10.1177/152574010935937.

Petersen, D. B., Chanthongthip, H., Ukrainetz, T. A., Spencer, T. D., & Steeve, R. W. (2017). Dynamic assessment of narratives: Efficient, accurate identification of language impairment in bilingual students. *Journal of Speech, Language, and Hearing Research, 60* (4), 983–998. 10.1044/2016_JSLHR-L-15-0426.

Petersen, D. B., & Spencer, T. D. (2012). The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education, 19* (4), 119–129. 10.1044/lle19.4.119.

Petersen, D. B., & Spencer, T. D. (2016). CUBED® Assessment. Laramie, WY: Language Dynamics Group, LLC https://www.languagedynamicsgroup.com/cubed/cubed-overview/.

Petersen, D. B., Spencer, T. D., Konishi, A., Sellars, T. P., Foster, M. E., & Robertson, D. (2020). Using parallel, narrative-based measures to examine the relationship between listening and reading comprehension: A pilot study. *Language, Speech, and Hearing Services in Schools, 51* (4), 1097–1111. 10.1044/2020_LSHSS-19-00036.

Petersen, D. B., Spencer, T. D., & Restrepo, M. A. (2016). *Spanish Narrative Language Measures Listening of the CUBED® Assessment*. Laramie, WY: Language Dynamics Group, LLC https://www.languagedynamicsgroup.com/cubed/cubed-overview/.

Petersen, D. B., & Stoddard, A. (2018). Psychometric requirements of oral and written language progress monitoring assessments. *Perspectives of the ASHA Special. Interest Groups, 3* (1), 180–197. 10.1044/persp3.SIG1.180.

Restrepo, M. A., Morgan, G. P., & Thompson, M. S. (2013). The efficacy of a vocabulary intervention for dual-language learners with language impairment. *Journal of Speech, Language, and Hearing Research, 56* (2), 748–765. 10.1044/1092-4388(2012/11-0173).

Silverman, R. D., McNeish, D., Speece, D. L., & Ritchey, K. D. (2019). Early screening for decoding-and language-related reading difficulties in first and third grades. *Assessment for Effective Intervention, 46*(2), 99-109. https://doi.org/10.1177/1534508419857234

Skarakis-Doyle, E., & Dempsey, L. (2008). Assessing story comprehension in preschool children. *Topics in Language Disorders, 28*(2), 131-148. https://doi.org/10.1097/01.tld.0000318934.54548.7f

Snow, C.E., Burns, M.S., & Griffin, P. (Eds.) (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press. https://doi.org/10.1002/pits.10011

Spencer, T. D., Moran, M., Thompson, M. S., Petersen, D. B., & Restrepo, M. A. (2020). Early efficacy of multitiered dual-language instruction: Promoting preschoolers' Span-ish and English oral language. *AERA Open, 6* (1). 10.1177/2F2332858419897886.

Spencer, T. D., & Petersen, D. B. (2018). Bridging oral and written language: An oral narrative language intervention study with writing outcomes. *Language, Speech, and Hearing Services in Schools, 49* (3), 569–581. 10.1044/2018_LSHSS-17- 0030.

Spencer, T. D., Petersen, D. B., Restrepo, M. A., Thompson, M., & Gutierrez Arvizu, M. N. (2019). The effect of Spanish and English narrative intervention on the language skills of young dual language learners. *Topics in Early Childhood Special Education, 38* (4), 204–219. 10.1177/0271121418779439.

Spencer, T. D., Petersen, D. B., Slocum, T. A., & Allen, M. M. (2015). Large group narrative intervention in Head Start preschools: Implications for response to intervention. *Journal of Early Childhood Research, 13* (2), 196–217. 10.1177/1476718X13515419.

Spencer, T. D., Weddle, S. A., Petersen, D. B., & Adams, J. L. (2017). Multi-tiered narrative intervention for preschoolers: A head start implementation study. *NHSA Dialog, 20* (1).

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. *New directions in discourse processing*, *2*(1979), 53-120.

Stothard, S. E., & Hulme, C. (1992). Reading comprehension difficulties in children. *Reading and Writing*, *4*(3), 245-256. https://doi.org/10.1007/bf01027150

Thordardottir, E. T., Weismer, S. E., & Smith, M. E. (1997). Vocabulary learning in bilingual and monolingual clinical intervention. *Child Language Teaching and Therapy*, *13*(3), 215-227. https://doi.org/10.1177/026565909701300301

Tomblin, J. B., & Zhang, X. (2006). The dimensionality of language ability in school-age children. https://doi.org/10.1044/1092-4388(2006/086)

Ukrainetz, T. A. (2006). The implications of RTI and EBP for SLPs: Commentary on LM Justice. *Language, Speech, and Hearing Services in Schools*, *37*(4), 298-303. https://doi.org/10.1044/0161-1461(2006/034)

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP). (2019). Reading Assessment. Retrieved from https://nces.ed.gov/nationsreportcard/reading/

Vaughn, S., Linan-Thompson, S., Mathes, P. G., Cirino, P. T., Carlson, C. D., Pollard-Durodola, S. D., ... & Francis, D. J. (2006). Effectiveness of Spanish intervention for first-grade English language learners at risk for reading difficulties. *Journal of Learning Disabilities, 39*(1), 56-73. https://doi.org/10.1177/00222194060390010601

Wackerle-Hollman, A., Spencer, T. D., Artman-Meeker, K., Kelley, E. S., Durán, L., & Foster, M. E. (2021). Multi-tiered system of supports in early childhood: identifying gaps, considerations for application, and solutions. *Early Childhood Research Quarterly, 56*, 201-212. https://doi.org/10.1016/j.ecresq.2021.03.010

Westby, C. (1985). Learning to talk - talking to learn: Oral-literate language differences. In C.S. Simon (Ed.), Communication Skills and Classroom Success: Therapy Methodologies for

Language-Learning Disabled Students. San Diego, CA: College Hill Press. https://doi.org/10.1177/026565908700300228

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical evaluation of language fundamentals—Preschool*, (CELF Preschool-2). Toronto, Canada: The Psychological Corporation.

Wiig, E., Secord, W. A., & Semel, E. (2009). *Clinical evaluation of language fundamentals preschool second edition–Spanish*. San Antonio, TX: The Psychological Corporation.

Table 1

Characteristics of the Samples Receiving English vs. Spanish Forms of NLM-Listening

English Forms Participants (*n*=65)		Spanish Forms Participants (*n*=61)	
Age in months, *M* (*SD*)	49 (6.7)	Age in months, *M* (SD)	52 (7.8)
Not reported	10	Not reported	16
Gender, *n* (%)		Gender, *n* (%)	
Male	22 (34)	Male	22 (36)
Female	22 (34)	Female	26 (43)
Not reported	21 (32)	Not reported	13 (21)
Race/Ethnicity, *n* (%)		Race/Ethnicity, *n* (%)	
White	10 (15)	White	1 (2)
Latino/Hispanic	18 (28)	Latino/Hispanic	49 (81)
Native American	28 (43)	Native American	0 (0)
Other	1 (2)	Other	1 (2)
Not reported	8 (12)	Not reported	10 (16)
Mother's Education, *n* (%)		Mother's Education, *n* (%)	
Elementary	2 (3)	Elementary	12 (20)
Some high school	5 (9)	Some high school	6 (10)
High school diploma	11 (17)	High school diploma	18 (29)
Some college	32 (49)	Some college	9 (15)
Bachelor's degree	7 (10)	Bachelor's degree	2 (3)
Master's degree	1 (2)	Master's degree	0 (0)
Not reported	7 (10)	Not reported	14 (23)

Annual Household Income, *n* (%)		Annual Household Income, *n* (%)	
less than $10,000	22 (34)	less than $10,000	22 (36)
$10,000-$21,999	20 (30)	$10,000-$21,999	21 (34)
$22,000-$29,999	5 (9)	$22,000-$29,999	5 (8)
$30,000-$39,999	10 (15)	$30,000-$39,999	2 (3)
$40,000-$49,999	3 (4)	$40,000-$49,999	1 (2)
$50,000 or more	0 (0)	$50,000 or more	0 (0)
Not reported	5 (9)	Not reported	10 (16)

Table 2

Descriptive Statistics for Spanish Retell Scores and Correlations with Two External Measures

Form	Valid *N*	Mean	*SD*	Median	Range	*r* with CELFtot	*r* with FWAYtot
Form 1	54	4.46	4.85	4.00	0-15	.58**	.61**
Form 2	59	6.81	6.03	7.00	0-21	.73**	.54**
Form 3	59	6.15	6.02	6.00	0-17	.71**	.62**
Form 4	57	5.65	5.39	5.00	0-19	.78**	.52**
Form 5	60	5.70	5.21	5.00	0-16	.70**	.58**
Form 6	57	5.54	5.69	4.00	0-17	.76**	.61**
Form 7	54	5.61	5.88	4.00	0-17	.78**	.53**
Form 8	53	5.94	5.95	4.00	0-17	.75**	.56**
Form 9	57	6.11	5.76	5.00	0-17	.70**	.70**
Form 10	54	5.37	5.36	4.00	0-17	.73**	.44**

Form 11	55	5.71	5.73	5.00	0-18	.63**	.67**
Form 12	59	5.36	5.10	5.00	0-19	.65**	.57**
Form 13	57	5.88	5.89	5.00	0-18	.70**	.57**
Form 14	57	6.46	6.24	5.00	0-19	.74**	.61**
Form 15	54	4.13	4.67	2.50	0-17	.63**	.48**
Form 16	57	6.18	6.11	6.00	0-19	.74**	.73**
Form 17	59	5.90	5.96	6.00	0-18	.72**	.63**
Form 18	59	5.24	5.32	4.00	0-18	.73**	.44**
Form 19	60	5.65	5.80	4.00	0-17	.75**	.61**
Form 20	57	4.82	5.31	4.00	0-15	.71**	.51**
Form 21	54	4.85	5.11	4.00	0-17	.74**	.61**
Form 22	57	5.95	5.91	5.00	0-18	.67**	.58**
Form 23	59	4.80	4.93	3.00	0-18	.69**	.52**
Form 24	57	6.39	6.00	6.00	0-19	.72**	.67**
Form 25	57	5.82	5.29	6.00	0-17	.71**	.61**

Note: CELFtot = total scores for the Clinical Evaluations of Language Fundamentals—Preschool in Spanish; FWAYtot = total word scores for *Frog, Where Are You?* retells. Total *N* = 61 for Spanish-language sample of participants.

**p < .005.

Table 3

Descriptive Statistics for English Retell Scores and Correlations with Two External Measures

Form	Valid *N*	Mean	*SD*	Median	Range	*r* with CELFtot	*r* with FWAYtot
Form 1	54	7.93	4.61	7.50	0-17	.33*	.27*
Form 2	59	7.85	4.39	8.00	0-19	.47**	.23
Form 3	59	8.51	5.07	9.00	0-21	.49**	.26*
Form 4	60	8.97	5.08	9.00	0-18	.50**	.31*
Form 5	58	7.21	4.55	7.00	0-18	.55**	.40**
Form 6	58	8.10	5.28	7.00	0-17	.54**	.26*
Form 7	56	6.89	4.31	6.00	0-15	.50**	.32*
Form 8	62	8.13	4.82	8.00	0-19	.40**	.17
Form 9	56	6.57	4.29	5.50	0-18	.50**	.21
Form 10	60	7.62	5.34	7.00	0-20	.55**	.38**
Form 11	58	6.88	4.07	7.00	0-18	.50**	.40**
Form 12	58	7.48	4.71	7.00	0-19	.30*	.30*
Form 13	58	7.60	4.83	7.00	0-19	.53**	.18
Form 14	59	7.47	4.43	7.00	0-18	.51**	.29*
Form 15	58	8.29	4.80	9.00	0-17	.40**	.11
Form 16	62	7.63	4.41	7.00	0-17	.38**	.35*
Form 17	58	6.55	4.21	6.00	0-17	.56**	.40**
Form 18	56	7.68	4.74	7.00	0-18	.46**	.31*
Form 19	57	7.44	4.68	7.00	0-17	.47**	.22
Form 20	59	8.46	4.74	8.00	0-19	.46**	.46**
Form 21	58	7.62	4.50	8.00	0-18	.53**	.37**
Form 22	60	8.00	4.36	7.00	0-17	.50**	.31*
Form 23	62	7.92	4.82	7.00	0-19	.37**	.28*
Form 24	59	7.37	4.57	8.00	0-21	.40**	.12

Form 25	58	7.38	5.16	6.00	0-18	.62**	.29*

Note: CELFtot = total scores for the Clinical Evaluations of Language Fundamentals—Preschool in Spanish; FWAYtot = total word scores for *Frog, Where Are You?* retells. Total *N* = 65 for English-language sample of participants.

***p* < .005.

Table 4

Evaluation of Congeneric, Essentially Tau-equivalent, and Essentially Parallel Forms Models for the Nine Benchmark Forms (1-9) of the Spanish and English Narrative Language Measures

	Model	MLR Scaled χ^2 (*df*)	RMSEA [90% CI]	CFI[a]	SRM R	Models Compare d	MLR Scaled $\Delta\chi^2$ (Δdf)[b]
			Spanish NLM-Listening Retells				
1	Congeneric (one-factor model, unconstrained)	18.892 (27)	.000 [.000, .052]	1.000	.020	--	--
2a	Tau equivalent (equal loadings)	31.923 (35)	.000 [.000, .081]	1.000	.088	2a vs. 1	14.988 (8)
2b	Tau equivalent (equal loadings except Form 1)	25.119 (34)	.000 [.000, .052]	1.000	.047	2b vs. 2a	6.296 (1)*
3a	Parallel (equal loadings; equal residual variances)	46.200 (43)	.035[.000, .095]	--	.082	3a vs. 2a	15.991 (8)*
3b	Parallel (2b plus equal residual	36.830 (41)	.000 [.000, .074]	--	.055	3b vs. 2b	13.287 (7)

	variances except Form 1)						
3c	Parallel (2b plus equal residual variances except Forms 1 and 2)	32.368 (40)	.000 [.000, .059]	--	.049	3c vs. 3b	5.463 (1)*
			English NLM-Listening Retells				
4	Congeneric (one-factor model, unconstrained)	46.248 (27)	.105 [.049, .155]	.943	.052	--	--
5	Tau equivalent (equal loadings)	58.067 (35)	.101 [.051, .145]	.931	.086	5 vs. 4	11.150 (8)
6	Parallel (equal loadings; equal residual variances)	66.817 (43)	.092 [.045, .134]	--	.100	6 vs. 5	8.440 (8)

Note. MLR = maximum likelihood estimator with robust standard errors and chi-square in Mplus; *df* = degrees of freedom; RMSEA = Root Mean Square Error of Approximation; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual; CI = Confidence Interval; *p < .05. [a]CFIs are not reported for models with constrained residual variances because the baseline model used in computing the CFI is not nested within this constrained model (see Widaman & Thompson, 2003). [b]The scaled chi-square difference test was used to test differences between nested models.

Table 5

Summary Statistics (across 25 Spanish Forms) of Correlations of Spanish NLM-Listening Retells with CELF-P and FWAY Scores

External Criterion Measure	Corrected mean $r^{a,b}$	SD_r	Median r	Minimum r	Maximum r
CELF-total	.71	.05	.72	.58	.78
CELF-WS	.59	.05	.60	.48	.66
CELF-SS	.74	.04	.74	.67	.82
CELF-EV	.59	.06	.60	.45	.68
CELF-FD	.67	.06	.68	.51	.74
FWAY-total	.59	.07	.58	.44	.73
FWAY-MLU	.46	.09	.46	.24	.67
FWAY-NDW	.63	.07	.62	.49	.74
FWAY-NTW	.58	.08	.57	.41	.73

Note: CELF-Total, CELF-WS, CELF-SS, CELF-EV and CELF-FD denote total scores, Word Structure scores, Sentence Structure scores, Expressive Vocabulary scores, and Following Directions scores for the Clinical Evaluations of Language Fundamentals—Preschool in Spanish; FWAYtotal, FWAY-MLU, FWAY-NDW and FWAY-NTW denote number of total words, mean length of utterance, number of different words, and use of subordination for *Frog, Where Are You?* retells in Spanish. [a]Pairwise sample sizes for correlations ranged from *N*=53-60 for CELF Measures and *N*=45-53 for FWAY measures. [b]Mean correlations were computed by applying Fisher’s *r* to *z*′ transformation, computing the sample size weighted mean of *z*′, and transforming the weighted mean of *z*′ back to *r*.

Table 6

Summary Statistics (across 25 English Forms) of Correlations of English NLM-Listening Retells with CELF-P and FWAY Scores

External Criterion Measure	Corrected mean $r^{a,b}$	SD_r	Median r	Minimum r	Maximum r
CELF-total	.48	.08	.50	.30	.62
CELF-WS	.41	.08	.41	.22	.55
CELF-SS	.47	.06	.47	.33	.62
CELF-EV	.45	.07	.45	.34	.62
CELF-FD	.53	.07	.52	.40	.76
FWAY-total	.29	.09	.29	.11	.46
FWAY-MLU	.28	.09	.28	.13	.47
FWAY-NDW	.45	.09	.45	.25	.56
FWAY-NTW	.42	.09	.42	.25	.54

Note: CELF-Total, CELF-WS, CELF-SS, CELF-EV and CELF-FD denote total scores, Word Structure scores, Sentence Structure scores, Expressive Vocabulary scores, and Following Directions scores for the Clinical Evaluations of Language Fundamentals—Preschool in English; FWAYtotal, FWAY-MLU, FWAY-NDW and FWAY-NTW denote number of total words, mean length of utterance, number of different words, and use of subordination for *Frog, Where Are You?* retells in English. [a]Pairwise sample sizes for correlations ranged from *N*=53-60 for CELF Measures and *N*=45-53 for FWAY measures. [b]Mean correlations were computed by applying Fisher’s *r* to *z*′ transformation, computing the sample size weighted mean of *z*′, and transforming the weighted mean of *z*′ back to *r*.

Figure 1

Distributions of 300 Alternate-form Correlations for Spanish and English NLM-Listening Retells

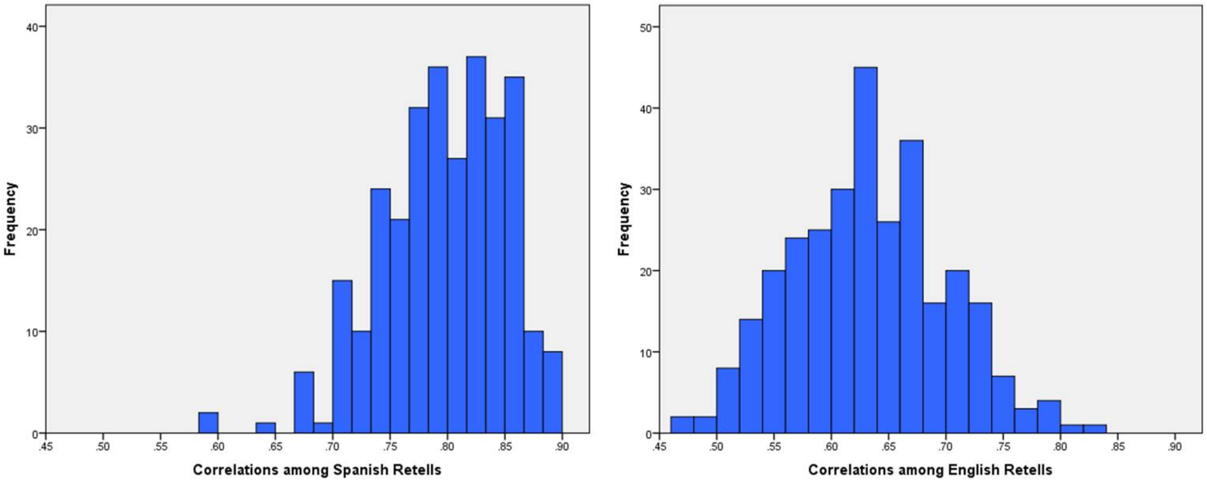
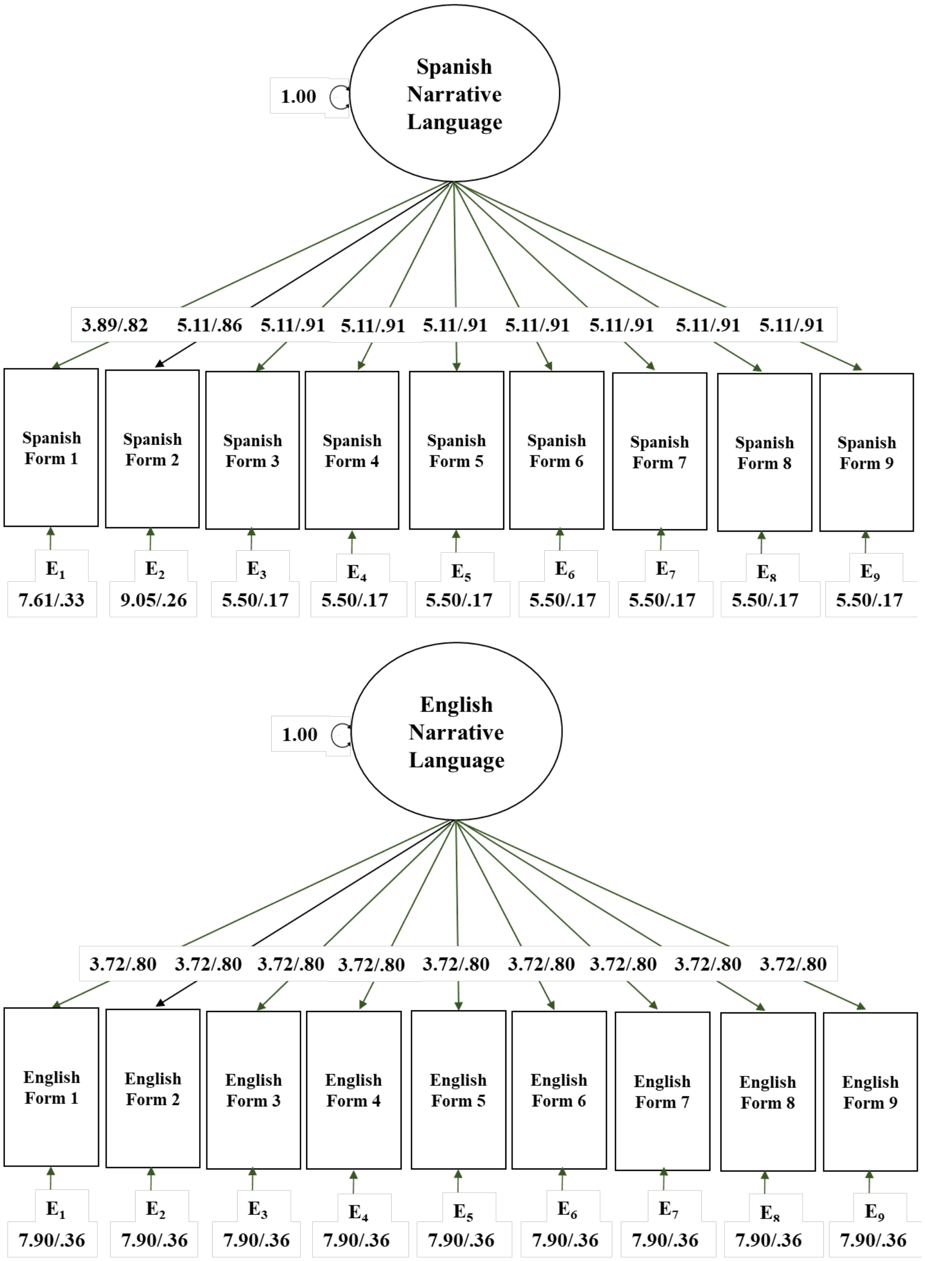


Figure 2

Unstandardized/standardized Parameter Estimates for Confirmatory Factor Analysis Tests for Parallel Forms



Note. The top panel is for the Spanish NLM-Listening benchmark forms (Model 3c on Table 3). The bottom panel is for the English NLM-Listening benchmark forms (Model 6 on Table 3). All parameters were statistically significant, $p \leq .001$.