2019-04-01

The Accuracy of a Spanish Dynamic Assessment of Narrative Language in Identifying Language Disorder: A Cross Validation Study

Mariah Forbush Romero
Brigham Young University

The Accuracy of a Spanish Dynamic Assessment of Narrative Language in Identifying Language Disorder: A Cross Validation Study

Mariah Forbush Romero

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Douglas B. Petersen, Chair
Kathryn Cabbage
Richard Sudweeks

Department of Communication Disorders

Brigham Young University

ABSTRACT

The Accuracy of a Spanish Dynamic Assessment of Narrative Language in Identifying Language Disorder: A Cross Validation Study

Mariah Forbush Romero
Department of Communication Disorders, BYU
Master of Science

This cross-validation study investigated the extent that a Spanish narrative language dynamic assessment accurately identified students with and without language disorder across three separate samples of bilingual and monolingual Spanish-speaking students from Guatemala, Mexico, and the U.S. Students with language disorder and students with typically developing language were administered a narrative dynamic assessment in Spanish. A test-teach-retest format of dynamic assessment was followed and student modifiability, or learning ability, was rated directly following the teaching phase of the assessment. Results indicated that the most predictive dynamic assessment variables for the Guatemalan sample were posttest scores combined with two separate modifiability measures (i.e., total modifiability scores and modifiability final judgment scores). These same variables were applied in the cross-validation classification analyses of the Mexico and U.S. samples with good classification accuracy achieved. The results of this study indicate that a Spanish narrative dynamic assessment may be a culturally appropriate diagnostic tool in identifying Spanish-speaking students with language disorder.

Keywords: dynamic assessment, language disorder, school-age children, narratives

ACKNOWLEDGMENTS

The author would like to thank the following committee members for their meaningful contribution: Douglas B. Petersen, Kathryn Cabbage, and Richard R. Sudweeks. Their expertise, guidance, and experience were vital to the success of this project. The author also thanks Allison Long and Maria Nelly Gutierrez as well as all members of the DYMOND team for their time and effort collecting and analyzing the data used in this project.

TABLE OF CONTENTS

LIST OF TABLES

DESCRIPTION OF THESIS STRUCTURE

This thesis, *The Accuracy of a Spanish Dynamic Assessment of Narrative Language in Identifying Language Disorder: A Cross Validation Study,* is written in a journal publication style format. The preliminary pages conform to university thesis requirements while the body of this thesis and its appendices follow standard APA formatting requirements. Portions of this manuscript may be adapted and submitted for publication in a peer-reviewed journal with the first author listed as a contributing author. Appendix A consists of a scoring rating scale used during the administration of the dynamic assessment. Appendix B consists of a parental consent form. Appendix C consists of an annotated bibliography.

Introduction

The Spanish-speaking population in the U.S. and internationally is growing at a rapid rate. In 2016, there were approximately 58.9 million Hispanics living in the U.S. constituting 17.6% of the total U.S. population (U.S. Census Bureau, 2018). From 2000 to 2010, the Hispanic population grew by 43%, accounting for 56% of the U.S. population growth (U.S. Census Bureau, 2011). Future projections indicate that this dramatic growth rate will only continue. Estimates from the U.S. Census Bureau suggest that between now and the year 2060, the Hispanic population in the U.S. will more than double, reaching 119 million (Vespa, Armstrong, & Medina, 2018).

Consequential to this national Hispanic population increase, the percentage of students coming from Hispanic homes has increased as well. From 1996 to 2016 the number of Hispanic students enrolled in a U.S. elementary school, middle school, high school, or college doubled from 8.8 million to 17.9 million (Bauman, 2017). The Hispanic population is rapidly increasing at an international level as well. According to the Cervantes Institute, Spanish is the second-most spoken language with almost 500 million native Spanish-speakers living throughout the world. By 2050, the number of native Spanish-speakers is predicted to rise to 754 million due an increasing Hispanic population in Spanish-speaking countries and in the U.S. (Cervantes Institute, 2016).

With a growing national Hispanic population and an increasing number of Hispanic children being introduced into schools worldwide, there will also be an increase in Spanish-speaking children with a language disorder. Current data show that approximately 7% of the U.S. school-age population has a language disorder (Tomblin et al., 1997). This prevalence of language disorder is expected to be similar in all school-age populations worldwide, including the Spanish-speaking population. Because of the increasing number of Hispanic students in schools throughout the world, there is a current and future need for SLPs to be prepared to validly assess this growing population of Spanish-speakers with language disorder (LD).

Current Practices Assessing Language Disorder in Spanish-Speakers

Currently, norm-referenced tests (NRTs) are the most common tool used to identify LD, especially in the U.S. However, most NRTs are ineffective in identifying LD in school-age children, yielding weak sensitivity and specificity (Spaulding, Plante, & Farinella, 2006). *Sensitivity* relates to how well a test correctly identifies children with a language disorder. *Specificity* relates to how well a test correctly identifies children without a language disorder. In order to meet standards of evidence of validity, diagnostic assessment measures should have sensitivity and specificity measures at or above 80% (Spaulding et al., 2006). Assessments that fail to meet these standards have a higher chance of misidentifying students with and without LD. Spaulding et al. (2006) reviewed 43 norm-referenced measures designed to identify LD. Of the 43 measures reviewed, only five NRTs reported sensitivity and specificity measures above 80%.

The classification accuracy of NRTs is often even weaker when used with culturally and linguistically diverse (CLD) students, often due to their inclusion of culturally biased test items, materials, and procedures. Laing and Kamhi (2003) claimed that English language NRTs are culturally inappropriate for use with CLD populations due to three main biases common to NRTs: (a) content bias, (b) linguistic bias, and (c) the disproportionate representation of ethnicities in normative samples. Content bias occurs when test items or testing procedures used to assess students assume that the life experiences (e.g., exposure to vocabulary, early literacy experience, and traditions of teacher to student interactions) of students are the same as the mainstream culture. Linguistic bias occurs when the language or dialect used by the child is different from the language expected of them by the examiner and by the test. Linguistic biases within NRTs are problematic in that they may incorrectly classify a child as atypical when the child may in fact have typical language development when viewed within the context of their mother language or dialect. NRTs are often further biased by the fact that there exists a disproportionate representation of ethnicities in normative samples. In the past, CLD students were often excluded from the normative groups used to establish norm-referenced assessment norms. Although most tests now incorporate students from various ethnicities into their sample data in an attempt to more fully represent the diversity present in U.S. schools, NRTs often continue to under-represent diverse students, and many CLD children with LD continue to be under-represented (Laing & Kamhi, 2003).

While research has indicated that many English-based NRTs may be biased measures for evaluating language status in CLD students, classifying these students is further complicated by the fact that NRTs designed specifically for Spanish-speaking students also often fail to meet acceptable standards of evidence of validity. For example, Restrepo and Silverman (2001) investigated the classification accuracy of the Spanish Preschool Language Scale (SPLS-3; Zimmerman, Steiner, & Pond, 1993) in identifying LD in Spanish-speaking students. They found that fifty-one percent of students from a local sample scored more than one standard deviation below the mean when compared to the normative data of Spanish speakers provided by the SPLS-3, and that many test items were considered culturally inappropriate. Additionally, the SPLS-3 was found to lack evidence of criterion-related validity when compared with other criterion-referenced measures of language ability.

The results of another study by Barragan, Castilla-Earls, Martinez-Nieto, Restrepo, and Gray (2018), found that the Clinical Evaluation of Language Fundamentals Fourth Edition - Spanish version (CELF-4S; Semel, Wiig, & Secord, 2006), the most commonly used Spanish language NRT, lacked high classification accuracy when used to identify LD in Spanish-bilingual elementary-school children from a low socioeconomic background. The researchers found that when they used the CELF-4S suggested score of 85 (-1.0 *SD*) to differentiate typical from disordered, 58% of students' scores fell below 1.0 standard deviation of the mean resulting in a sensitivity of 93% and specificity of 65%. The results from both Restrepo and Silverman (2001) and Barragan et al. (2018) indicate that even NRTs designed for Spanish-speakers may lack high sensitivity and specificity, and may result in an over and under classification of Spanish-speaking children with LD.

In addition to NRTs, criterion-referenced tests (CRTs), also known as domain-referenced tests, are commonly used to identify LD in school-age students. CRTs differ from NRTs in that rather than compare an individual's performance to same-age peers, CRTs compare a student's performance on a language task with a predetermined standard of performance, often based on developmental data (Popham & Husek, 1969). While CRTs can provide valuable information regarding a student's ability to master or achieve specific tasks, and can be used to monitor a student's progress, when used with CLD populations, criterion-referenced measures often have the same weaknesses that NRTs do because of their static nature (Dziuban & Vickery, 1973). Criterion-referenced tests and NRTs traditionally only provide information on a student's current knowledge and performance. This static information is often confounded by diversity in cultural and linguistic experience making it difficult to differentiate language difference from disorder.

In addition to the challenges identifying Spanish speakers in the U.S., using static measures for the identification of LD in Spanish-speaking countries is also questionable due to a limited number of assessments designed specifically for native Spanish speakers. For example, Spanish language NRTs that are normed using the U.S. bilingual Spanish-speaking population may not be representative of Spanish-speaking populations from other countries. Furthermore, static Spanish criterion-referenced tests cannot assess the underlying construct of language learning ability and can therefore yield results that are confounded by prior experience and language exposure. More accurate and culturally appropriate assessment measures are needed to identify LD in Spanish-speaking children in the U.S. and abroad.

Dynamic Assessment

Dynamic assessment is a promising alternative to NRTs. Dynamic assessments differ from NRTs in that they are dynamic measures of student learning rather than single static measures of a student's knowledge, or ability, at a given point in time. Dynamic assessment draws on Vygotsky's Zone of Proximal Development (ZPD; Vygotsky, 1978) principles by measuring student learning, often using a pretest-teach-retest model. A child's zone of proximal development falls between what tasks the child can successfully perform independently and what tasks are completely outside their ability to accomplish at their current level of functioning. A child's ZPD is what a child can do when given appropriate instruction and support.

By using a pretest-teach-retest model, dynamic assessment allows an examiner to determine, during pretesting, how well a student can perform tasks independently. Using this information, the student's ZPD is then investigated during the teaching phase, where Mediated Learning Experiences (MLE; Feuerstein, Rand, & Hoffman, 1979) are provided. It is during this teaching phase that the examiner provides individualized instruction to determine a student's learning potential, or modifiability. *Modifiability* is a measure of examiner effort and child responsivity. Examiner effort refers to how much effort is required from the test examiner to help a child learn and make progress within MLE sessions. Child responsivity, as defined by Peña et al. (2006), is "changes in children's cognitive strategies within the MLE sessions" (p. 1039). In other words, throughout the teaching phase, as the examiner works from a high level of support to a low level by slowly reducing the amount of instructional supports given to the student, they take note of how well the child can apply cognitive learning strategies to the language task. The ability a child has to continue to apply these cognitive learning strategies when given less support is an indicator that learning has occurred. A child's ability or inability to apply cognitive strategies to complete a linguistic task, and the effort required from an examiner to help a child do so, are indicators of how difficult it is for that particular student to learn language.

In this respect, modifiability outcomes from dynamic assessment are unique in that their purpose is to not compare a student's performance to their peers using norm referencing. Rather, measuring modifiability allow examiners to determine how able a child can use cognitive processes in a learning environment. In doing so, dynamic assessment avoids much of the cultural biases present in NRTs and criterion-referenced assessments by providing insight into the cognitive, language abilities of students rather than measuring their current language competence, which could be confounded by multiple variables, including limited English language proficiency, poor normative sample representations, socioeconomic status, or other cultural differences.

Because of the potential benefit for using dynamic assessments with CLD populations, and because dynamic assessments may be valuable alternatives to NRTs and criterion-referenced tests, several studies have investigated the role of dynamic assessments of language in classifying LD. For example, Kapantzoglou, Restrepo, and Thompson (2012), Peña, Quinn, and Iglesias (1992), and Ukrainetz, Stacey, Walsh, and Coyle (2000) investigated the classification accuracy of various dynamic assessments of vocabulary. Additionally, Petersen and Gillam (2015) studied the role of dynamic assessment of reading in identifying children with reading disorders. However, narrative language has been the primary focus of dynamic assessment of language research, as narratives are highly effective measures of language ability.

Narrative Language

Narratives are replete with academic language, containing complex grammatical forms and vocabulary. This complexity of academic language often makes it difficult for children with LD to comprehend narrative language receptively and use it expressively. For this reason, difficulty understanding and producing narrative language can be a key indicator of LD (Boudreau, 2008). For example, in a study designed to determine the differences in grammatical accuracy and quality of content in the narratives of students with LD when compared to the narratives of typically developing students, Colozzo, Gillam, Wood, Schnell, and Johnston (2011) found that students with LD scored significantly lower in narrative form and content when compared to a typically developing group of students. Furthermore, participants with LD tended to have stronger content in their stories than grammar or vice versa. Rarely were the participants’ grammatical scores and content scores matched in complexity.

In a longitudinal study of bilingual Spanish-English speaking children, Squires et al. (2014), sought to determine whether or not bilingual children with LD differ from their typically developing peers in macro and micro narrative structure from kindergarten to first grade. Macrostructure refers to the content of a narrative such as the inclusion of a character, problem, consequence, etc. Microstructure refers to the grammatical structure of a narrative. The results of the study determined that in terms of macrostructure and microstructure, the typically developing group had overall higher scores than the LD group. Additionally, while both groups improved in macrostructure narrative abilities between kindergarten and first grade, the typically developing group made significantly larger gains than the LD group. In terms of microstructure, the typically developing students made gains over time, but no change was observable in the LD group.

The results of these studies indicate that the ease with which a child can learn the macro- and microstructural, or content and grammatical elements of narrative language may be promising indicators to identifying LD in both monolingual and bilingual students. Because of the connection between narrative language and LD, researchers have sought to develop increasingly valid and reliable dynamic assessments of narrative language that can be effectively used to identify CLD students with LD.

Dynamic Assessment of Narrative Language

An increasing amount of research in the last 20 years has indicated the benefit of using narrative dynamic assessments to separate difference from disorder in CLD student populations. In 2001, Miller, Gillam, and Peña developed an English narrative language assessment, called the Dynamic Assessment Instrument (DAI; Miller, Gillam, & Peña, 2001), which was designed to identify CLD students with LD. In 2006, Peña et al., administered the assessment to 71 English-speaking first and second graders from Latino American, African American, and European American ethnic backgrounds. The dynamic assessment consisted of a pretest phase in which students produced an oral narrative using a wordless picture book. Following the pretest phase, the typically developing and LD groups participated in two 30-minute mediated learning experiences (MLE) in which narrative components were taught. Two 5-point rating scales were used to score each child’s overall modifiability following the MLE sessions—the first scale rated examiner effort while the other rated child responsivity. Lastly, during the posttest phase, participants produced a final narrative using a different wordless picture book. The researchers were particularly interested in determining which combination of dynamic assessment scores (pretest, posttest, and/or modifiability scores) had the highest classification accuracy in identifying LD amongst the participants. Using discriminant function analyses, Peña et al. (2006) found modifiability rating scores, based on examiner effort and child responsivity, to have the highest sensitivity (93%) and specificity (96%) when compared to the sensitivity and specificity of the pretest or posttest scores. However, when posttest scores and modifiability scores were combined, classification accuracy increased to 100%.

In a replication study, Kramer, Mallet, Schneider, and Hayward (2009) investigated the classification accuracy of the same narrative Dynamic Assessment Instrument (DAI; Miller et al., 2001) examined by Peña et al. (2006) and Peña, Gillam, and Bedore (2014) when administered to 17 (12 students with typical language development, 5 students with disordered language development) third-grade First Nation students living on the Samson Cree Nation Reserve in Alberta, Canada. The third graders were bilingual English-Cree speaking students, receiving classroom instruction in both languages. The administration of the dynamic assessment matched the administration procedures as detailed in Miller et al. (2001), Peña et al. (2006), and Peña et al. (2014). The same wordless picture books were used as well as the same scoring Likert scales to measure and score the students' modifiability and narrative productions. Similar to Peña et al. (2006), the results of the study also reported a high classification accuracy. While both the typical language learning students and the atypical language-learning students had similar pretest scores, typical language learning students made greater improvements in targeted and non-targeted narrative elements. A combination of modifiability and gain scores accurately classified students, yielding 100% sensitivity and 92% specificity. Gain scores alone also resulted in the same degree of sensitivity and specificity.

In 2014, Peña et al. used a research design similar to their 2006 study to determine if an English narrative dynamic assessment would accurately identify LD specifically in Spanish-English bilingual speakers. The research participants consisted of 54 bilingual, kindergarten children (18 with LD, 36 with typical language development). Eighteen of the students with typical language were matched according to age, sex, language experience, and IQ with the 18 students with LD. Prior to the administration of the dynamic assessment, each student first participated in a language sample. The dynamic assessment followed a pretest-teach-posttest format including two 30-minute MLE sessions and was administered across two separate days. The main focus of the MLE sessions was to help students learn to tell complete narratives. During the first MLE session, instruction was focused on the main story grammar elements (e.g., character, setting, initiating event, problem, consequence, etc.) missing in the student's pretest narrative. During the second MLE session, the clinician and student co-created a narrative. After each MLE, the clinician rated the child's modifiability and scored the child's stories based on the number of story components included, use of episode structure, and complexity of language. The results of the study indicated that a combination of modifiability, posttest, and language sample scores analyzed using the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) software yielded the highest classification accuracy (sensitivity: 89-100%, specificity: 89%). Peña et al. (2014) concluded that English dynamic assessments are accurate in identifying LD in bilingual, Spanish-English speakers.

Drawing on Peña et al.'s 2006 and 2014 foundational research, Petersen, Chanthongthip, Ukrainetz, Spencer, and Steeve (2017) also investigated the accuracy of an English narrative dynamic assessment when used with bilingual Spanish-English speakers, including more dominant Spanish-speaking bilinguals. In doing so, Petersen et al. (2017) sought to identify the best combination of scores (modifiability, gain, posttest, etc.) that yielded the highest classification accuracy of LD in Spanish-English speakers. Additionally, as previous dynamic assessments were somewhat lengthy, often requiring multiple days to complete, the researchers sought to design a more efficient and condensed version of narrative dynamic assessment by utilizing real time scoring and by reducing the entire length of administration of two sessions (each session consisting of pretesting, teaching, and posttesting) to 25 minutes each, without losing classification accuracy. Similar to Peña et al. (2006), and Peña et al. (2014), Petersen et al. (2017) found the modifiability ratings from both dynamic assessment sessions to yield a high classification accuracy (100% specificity and 100% sensitivity). Additionally, a modifiability rating from only one 25-minute session, when combined with the duration score, yielded 100% sensitivity and 91% specificity, indicating that it may be possible to further shorten the administration length of the narrative dynamic assessment to make it more clinically useful. Due to these findings, Petersen et al. (2017) concluded that a shortened version of an English narrative dynamic assessment was accurate in identifying LD in bilingual, Spanish-English speakers.

The Current Study

The current study was designed to examine the classification accuracy of a Spanish narrative dynamic assessment administered to monolingual Spanish-speaking Guatemalan students, and to cross-validate that dynamic assessment with an independent sample of monolingual Spanish-speaking students from Mexico, and an independent sample of bilingual English/Spanish-speaking typically developing students from the U.S. There is currently no narrative dynamic assessment that has adequate evidence of validity designed specifically for use with monolingual Spanish speakers. The need for a more valid means of identifying Spanish-speaking students with a LD is more and more pressing each year as the Spanish-speaking population increases worldwide and across the U.S. Identifying a Spanish narrative dynamic assessment that has strong evidence of validity could be of significant clinical use, especially if the classification accuracy of the dynamic assessment is higher than commonly-used NRTs. Although gain scores and duration of the teaching phase have been shown to aid in the identification of language ability in some studies, consistently the highest predictive evidence of validity is found when modifiability measures are added to posttest scores. Therefore, we hypothesize that the strongest classification accuracy across multiple Spanish-speaking samples of students will be achieved when modifiability scores are added to the posttest score of a Spanish dynamic assessment.

In order to test this hypothesis, the following research questions were explored:

1. How much variance in language ability (R^2) do Spanish dynamic assessment modifiability variables, when added to the posttest dynamic assessment variable, account for in Guatemalan, monolingual Spanish-speaking school-age students? Do dynamic assessment gain scores and the teaching phase duration account for variance in language ability over and above the dynamic assessment alone?
2. What is the optimal sensitivity and specificity of the Spanish dynamic assessment for monolingual Spanish-speaking students from Guatemala with and without language disorder?
3. What clinically interpretable cut-points provide the highest sensitivity and specificity?
4. Using the specific combination of dynamic assessment variables and cut-points for those dynamic assessment measures identified with students from Guatemala, how well does

the Spanish dynamic assessment of language account for language variance and correctly classify a cross-validated sample of monolingual Spanish-speaking students from Mexico with and without LD and another cross-validated sample of bilingual Spanish-speaking students from the U.S. who are typically developing?

5. Using the specific combination of dynamic assessment variables and cut-points for those dynamic assessment measures identified with students from Guatemala, how well does the Spanish dynamic assessment of language account for language variance and correctly classify all participants combined?

Method

Participants

The participants included a total of 59 Spanish-speaking school-age children with typically developing (TD) language and 9 school-age children with language disorder (LD). Of those participants, 28 (LD = 4, TD = 24) were monolingual Spanish-speaking students from Guatemala, 20 (LD = 5, TD = 15) were monolingual Spanish-speaking students from Mexico, and 20 (TD = 20) were bilingual English/Spanish-speaking students from the U.S (Table 1 and Table 2). Students from Mexico and the U.S. were included as fully independent, cross-validated groups and were not matched in any way to the students from Guatemala. To protect the human participants involved in this study, approval was obtained from the Institutional Review Board at Brigham Young University prior to the recruitment of all participants in the study. Additionally, prior to the administration of any assessments, parent consent forms were obtained from child guardians and the teachers who completed the forms. Any students whose form was not completed did not participant in the study.

Participants from Guatemala were recruited from a school consisting of 47 school-age children ranging in age from 4 to 18. The school was a private school which focused on integrating a high percentage of students with special needs into general education classrooms. Examiners contacted the school director and teachers to request permission to administer the dynamic assessment. Of the 47 students enrolled, 32 students were ultimately evaluated by the examiners. However, three of the students were excluded from the study due to age, and one student was excluded from the study because he had Down syndrome. Of the final 28 participants, 10 were female and 18 were male with the average grade level being eighth grade.

Participants from Mexico were recruited from a public school by a collaborating researcher in the field of psychology. First through sixth grade students were enrolled in the school with both general education and special education instruction present. The participants from this region of Mexico came from largely low to mid-income level homes and had been attending school regularly from preschool. Parental consent forms in Spanish were sent home before participants were administered any assessments. From the parental consent forms returned, 20 students were selected as participants based on language status, grade level, and gender to obtain a more balanced sample. Of the 20 participants evaluated, 13 were male and 7 were female. Information regarding the grade level of students was gathered with the average grade level of the sample being Grade 2.

Participants from the U.S. were recruited from a larger sample of students participating in a large-scale diagnostic study. These bilingual participants all attended an elementary school in the Mountain West. Parental consent forms in Spanish and English were sent home to parents prior to the administration of the dynamic assessment. The primary language status of each child (monolingual Spanish, dominant Spanish with some English, or balanced bilingual

Spanish/English) was determined by consulting with SLPs, teachers, parents, and the students themselves. From this larger bilingual sample, 20 participants were selected. Of the 20 total participants, 11 were female and 9 were male with the average grade level being fifth grade. Details regarding participant demographics from all three samples can be found in Table 1 and Table 2.

Method for Determining Language Disorder

Researchers collaborated with partnering educators and researchers across the different data collection sites to identify participants with LD and similar participants with typical language prior to the administration of the dynamic assessment. For a student to be classified as

Table 1

Participant Demographic Information by Gender

Sample	Language Ability	Male	Female	Total
Guatemala	LD	2	2	4
	TD	16	8	24
Mexico	LD	4	1	5
	TD	9	6	15
U.S.	LD	0	0	0
	TD	9	11	20
Combined	LD	6	3	9
	TD	34	25	59

Note. LD = language disorder. TD = typically developing language.

Table 2

Participant Demographic Information by Grade Level Groups

Sample	K-2nd	3rd-5th	6-8th	9-10th	11-12th	Total
Guatemala	3	6	8	9	2	28
Mexico	9	11	0	0	0	20
U.S.	4	14	2	0	0	20
Combined	16	31	10	9	2	68

a child with LD they were required to meet three criteria. First, a special educator had to confirm that the student had a LD. Second, a student had to be receiving language services. Third, a student had to score at least one standard deviation below the mean on at least two indicators: a norm-referenced test in Spanish (e.g., CELF-4:Spanish), a language sample from the NLM Listening, a nonword repetition task, or two of four measures [mean length of utterance (MLU), total number of words (TNW), or number of different words (NDW)] from a language sample using the wordless picture book *Frog, Where Are You?* (Mayer, 1969). For a student to be classified as TD, they were required to meet three criteria. First, a special educator had to confirm that the student did not have a LD. Second, a student could not be receiving language services. Third, a student could not score more than one standard deviation below the mean on a norm-referenced test in Spanish (e.g., CELF-4: Spanish), a language sample from the NLM Listening, a nonword repetition task, or any of the four measures [mean length of utterance (MLU), total number of words (TNW), or number of different words (NDW)] from a language sample using the wordless picture book *Frog, Where Are You?* (Mayer, 1969).

Measures

The *Frog, Where Are You?* language sample. The *Frog, Where Are You?* wordless picture book was used to elicit a narrative language sample from the Guatemalan students as part of the criteria in determining LD. The elicitation of the *Frog, Where Are You?* narrative sample occurred prior to the administration of the dynamic assessment and took approximately seven minutes. First, the examiner showed the student a wordless picture book while reading a script. The examiner then gave the student the book and asked them to retell the story. All Frog Retell samples were audio recorded for language sample transcription and administration fidelity purposes. Each narrative language sample was later analyzed using TNW, NDW, and MLU. TNW, NDW, and MLU data was compared to the Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) database. The SALT Reference Database is composed of thousands of language samples taken using the *Frog, Where Are You?* picture book and includes a large number of bilingual Spanish/English and monolingual Spanish samples.

The CUBED: Narrative Language Measures (NLM). Similar to the role of the *Frog, Where Are You?* sample in determining LD in the Guatemalan sample, the NLM subtest of the CUBED assessment (Petersen & Spencer, 2016) was used to help determine LD in the Mexican and U.S. samples. The NLM is a progress monitoring tool of language and reading that includes brief, one-minute narrative retell measures. One NLM Spanish story was preselected and then used to elicit a brief narrative retell language sample from each student. The examiner first read the story to the student. Next, the examiner asked the student to retell the story. The student's retell was audio-recorded and scored in real-time. The student's raw score was then computed into a standardized score and compared to norms of 136 Spanish-speaking elementary school children. If the student's score fell below one-standard deviation of the mean it was an indicator of LD.

The Clinical Evaluation of Language Fundamentals, 4th Edition (CELF-4). Two subtests from the CELF-4 Spanish version were administered to the participants recruited from Mexico—the Concepts and Following Directions subtest and the Recalling Sentences subtest - as part of the criteria for LD (CELF-4S; Semel et al., 2006). Administration of both subtests took approximately 10 minutes to complete. Examiners followed the guidelines and script described in the CELF-4 examiner's manual to administer both subtests. The CELF-4 protocol was used to assess and score each student's performance on each subtest. The recorded scores were later compared to the normative data provided in the examiner's manual. Those students whose scores across both subtests were at or below one standard deviation below the mean were noted.

Nonword repetition task. A nonword repetition task was also used as a measure of LD with the Mexico and U.S. samples. Two researcher-generated nonwords were added to a sample of 10 nonwords from the Children's Test of Nonword Repetition (CNRep; Gathercole & Baddeley, 1996). Specifically, one 2-syllable low complexity nonword was first presented to each student (see Archibald & Gathercole, 2006, Appendix B, 2-syllable low complexity nonword #2), followed by two 3-syllable low complexity nonwords (see Archibald & Gathercole, 2006, Appendix B, 3-syllable low complexity nonwords #2 and #3), then a high complexity 4-syllable nonword (see Archibald & Gathercole, 2006, Appendix B, 4-syllable high complexity nonword #3), a low-complexity 4-syllable nonword (see Archibald & Gathercole, 2006, Appendix B, low complexity 4-syllable nonword #3), and a high-complexity 4-syllable nonword (see Archibald & Gathercole, 2006, Appendix B, 4-syllable high complexity nonword #1). The nonword test increased in difficulty thereafter, with the presentation of a low word-like

5-syllable nonword (see Archibald & Gathercole, 2006, Appendix A, low-word like nonword #14), two high word-like 4-syllable nonwords (see Archibald & Gathercole, 2006, Appendix A, high-word like nonwords #9 and #14), and a 5-syllable low word-like nonword (see Archibald & Gathercole, 2006, Appendix A, low-word like nonword #5). Finally, a 6-syllable researcher-generated nonword ‘pristorakitional’, and a high-word like researcher-generated 4-syllable nonword ‘ruderpation’ were presented. The 12 nonwords were audio-recorded by a bilingual speech-language pathologist who pronounced the words using Spanish language conventions whenever possible. The students were instructed to listen to the audio-recording and repeat back each word. The students’ responses were audio-recorded and later scored according to the number of correct syllables the students produced. The number of correct syllables was then totaled with the highest possible score being 51. Students’ scores who fell at or below 70% syllables correct (a score of 36 or lower) on the nonword repetition task were noted and this information was used to help determine the likelihood of the student having LD.

The Spanish narrative dynamic assessment. Three separate teams of examiners composed of trained, Spanish-speakers administered the dynamic assessment across the three separate locations—four examiners in Guatemala, two examiners in Mexico, and five examiners in the U.S. The administration of the dynamic assessment was conducted entirely in Spanish and consisted of three phases: a pretest phase, a teaching phase, and a posttest phase. During both the pretest and posttest phases, the student was read a story by the examiner and was then asked to retell that story. During the teaching phase, the story from the pretest phase was used to help the student learn and include story grammar elements into their narrative productions and to increase the linguistic complexity of their retells. Immediately after the teaching phase of the dynamic assessment, the examiner completed a modifiability rating form. This form required the examiner to reflect on their teaching interaction with the student, and to rate the student's language learning potential by assigning the student both a total modifiability score and a modifiability final judgment score. After the modifiability rating scale was completed, which took less than one minute, the posttest of the dynamic assessment was administered. All scoring was done in real time and occurred after each of the three phases, before the examiner moved on to the next phase. Total administration time of the dynamic assessment took approximately 10-15 minutes per student.

Dynamic assessment pretest and posttest. Two different stories were used during the pretest and posttest. Both the pretest and posttest stories were matched in linguistic complexity (parallel in story length, use of tier-two words, dual-episode story structure, inclusion of subordinate clauses). During the pretest and posttest phases, each examiner read the student a short story in Spanish. After the story was read to the student, the examiner asked the student to retell the story. As the student produced their narrative retell, the examiner scored their narrative in real-time using a point system. Each retell was scored out of 35 possible points. Two points were given for the inclusion of each of 13 story grammar elements for a total of 26 possible points. One point was also given per use of *because, when, after,* or *then* for a total of nine points. Following the pretest, the teaching phase was completed followed by the posttest. The posttest phase followed the same format as the pretest phase.

Teaching phase. The teaching phase, which lasted approximately five minutes, consisted of three main steps designed to help the student independently produce complete narrative episodes of greater linguistic complexity. First, the examiner retold the same story used during the pretest phase while simultaneously pointing to pictures and icons representing story grammar elements. The examiner took care to point out each icon while modeling the corresponding story grammar element (e.g., “This is the character. The character in the story is Sam.”). In the second step, the pictures and icons were placed in front of the student and the student then practiced retelling the same story while referring to the pictures and icons. During this portion of the teaching phase, the examiner focused their teaching on the story grammar elements the student omitted during the pretest or that they skipped during the teaching phase retell. If the student skipped a story grammar element, the examiner stopped the student’s retell and prompted them with an open-ended question (e.g., “Wait. Where was Sam?”) or provided a clear model (e.g., “Sam is in the kitchen. Say it like this, ‘Sam was in the kitchen’”). The examiner then encouraged the student to start at the previous story grammar element and to make sure to include the element they missed. The examiner may have also encouraged the student to include the adverbial subordinating conjunctions *because*, *when*, or *after* to increase the linguistic complexity of the student’s narrative retell. During the third step, the pictures were removed, and the student retold the story using only the icons. At this time, the examiner followed the same teaching procedures detailed in step two.

Modifiability. Directly following the teaching phase, the examiner rated the student’s modifiability using the modifiability rating form. This modifiability rating form is a modified version of the modifiability scoring sheet from the Predictive Early Assessment of Reading and Language (PEARL; Petersen & Spencer, 2014). Using a 5-point Likert scale, the examiner rated each student’s modifiability on six subcategories. The six subcategories are as follows: 1) response to prompts; 2) degree of transfer; 3) attention to teaching; 4) little to no frustration; and 5) disruptions. A rating of 4 indicated the student’s relative ease to learn, while a rating of 0 indicated that learning was difficult for the student. To achieve the total modifiability score, the examiner totaled each of the six subcategory ratings with a maximum score of 24. The examiner then assigned a modifiability final judgment rating score to the student to represent their overall clinical impression of the student's responsiveness to intervention. Both the total modifiability rating and the modifiability final judgment scores were used for data analysis.

Administrator Training

A Spanish-English bilingual, certified SLP and a native Spanish speaking researcher in Mexico who has her Ph.D. in psychology trained seven bilingual research assistants to administer the dynamic assessment, CELF-4S, nonword repetition task, and to collect language samples using the Frog retell and NLM. The research assistants practiced administering and scoring the dynamic assessment in real time with their peers. The trainers observed these practice sessions and provided feedback to reduce the number of examiner errors made and increase the standardization of administration. To insure administration fidelity, each research assistant was given a script to follow during the pretest and posttest phases of the dynamic assessment. Although the research assistants followed a script during the pretest and posttest phases, during the teaching phase they were encouraged to be flexible in responding to the specific needs of each individual student, while simultaneously adhering to the dynamic assessment teaching guidelines.

All administrations of the NLM, nonword repetition task, and dynamic assessment were audio-recorded for fidelity purposes. Administration of the dynamic assessment was also video recorded whenever possible, with the student being administered the test out of the video frame. Administration fidelity of each of these measures was examined using a fidelity checklist. Twenty-five percent of administrations were randomly selected and reviewed using each checklist. The administration fidelity of the NLM and nonword repetition were 100% and 87%, respectively. Fifteen (25%) of the dynamic assessment audio files were also selected and scored—five files from the Guatemala sample, five files from the Mexico sample, and five files from the U.S. sample. One-hundred percent administration fidelity was achieved with the Guatemala sample, 85% administration fidelity was achieved with the Mexico sample, and 92% administration fidelity was achieved with the U.S. sample. Across all 15 dynamic assessment audio files reviewed, an average of 92% administration fidelity was observed.

Interrater Reliability

Point-to-point interrater reliability of scoring for the Frog retell, NLM, nonword repetition task, and dynamic assessment was calculated from 25% of the sample (Table 3). Bilingual research assistants, who were blind to the participants' language abilities and were not the primary examiners, rescored each of the randomly selected audio-files.

For the Frog retell, transcription and C-unit segmentation agreement was determined using 25% of the Frog audio recordings from the Guatemalan sample. Mean point-to-point agreement was 89% for transcription and 92% for C-unit segmentation. Coding of the Frog retell was completed using the SALT software making it unnecessary to calculate scoring agreement. Interrater reliability for scoring of the NLM was estimated using 10 (25%) NLM audio recordings. Five recordings from the Mexico sample and five from the U.S. sample were randomly selected. Mean point-to-point agreement for this measure was 83% across both the Mexico and U.S. samples (78% for the Mexico sample and 87% for the U.S. sample) with a range of 56%-100%. Point-to-point interrater reliability of scoring for the nonword repetition task was also completed using 10 (25%) of the nonword repetition tasks from the Mexico and U.S. samples. Mean point-to-point agreement in scoring for this measure was 89% across all 10 selected participants (88% scoring reliability for Mexico and 89% scoring reliability for the U.S.) with a range of 84%-100%.

Fifteen (25%) of dynamic assessment pretests and posttests, as well as modifiability ratings, were checked for reliability. Across all 15 audio-recordings, scoring reliability of the pre- and posttest was 83% with a range of 69%-100%. Mean point-to-point scoring reliability of the modifiability rating was 82% with a range of 29%-100%. Two outlier scores significantly influenced these results. One selected audio recording from the Mexico sample resulted in mean interrater agreement of 29%, however, all other mean point-to-point agreement scorings in the Mexico sample were 71% or 100%. Additionally, one selected audio recording from the U.S. sample resulted in a mean interrater agreement of 57%, however, all other agreement scores were 86% or 100%. Recognizing that the purpose of the modifiability rating scale is to help clinicians differentiate typical from disordered, exact point-to-point agreement across modifiability scoring was not expected. When considering interrater agreement with dynamic assessment modifiability measures, consistency in general modifiability scoring is more valuable. For example, a student scoring at the low end of the rating scale (i.e., 0 and 1 points) is indicative of possible disorder while a student scoring in the average to high end (i.e., 2-4 points) is indicative of average to typical language. For this reason, two additional measures of interrater reliability in the scoring of the modifiability rating scale were calculated: First, modifiability scores within one point of each other were considered to be in agreement; and second, when the modifiability final judgement score was dichotomized with scores from 0-1 and scores from 2-4 considered to be in agreement. When these additional interrater modifiability calculations were completed, mean reliability for each individual sample, and across all samples combined, increased to 100%. Specific information regarding the scoring reliability of the dynamic assessment for each sample can be found in Table 3.

Data Analysis

Data were analyzed using the Statistical Package for Social Sciences (SPSS version 24.0; IBM Corp., 2016). Logistic regression and receiver operator characteristic (ROC) analyses were conducted to estimate how much variance in language ability was accounted for, determine which dynamic assessment variables were predictive, and to find the best combination of sensitivity and specificity for each sample group individually (i.e., Guatemala sample, Mexico sample, and U.S. sample) and when all three sample groups were combined. Logistic regression uses independent and continuous predictor variables to predict a binary dependent variable. In this study, language ability was used as the binary dependent variable (language disorder/no language disorder) while the predictor variables from the dynamic assessment acted as the independent variables. These same variables were used in the ROC analyses. ROC analyses, yielding area under the curve (AUC) results were conducted to identify optimal cut-points for the dynamic assessment, yielding the best balance of sensitivity and specificity (i.e., sensitivity at or above 80% with specificity as close to 80% or higher whenever possible). In addition to the use of logistic regression and ROC analyses, clinically interpretable cut-points were applied to each sample to find optimal sensitivity and specificity percentages.

Results

Guatemalan Sample

Hierarchical logistic regression was used to determine how much variance the Spanish dynamic assessment modifiability variables (total modifiability and modifiability final judgment) accounted for language ability when combined with the dynamic assessment posttest score (Question 1a). In the first hierarchical logistic regression model, the posttest variable was entered into the logistic regression first followed by the total modifiability score. Results of Model 1

Table 3

Interrater Reliability Results

Sample		DA Pre and Post	DA Mod	DA Mod +/- 1	DA Mod Di	NLM	NWRT
Guatemala	Mean	95%	86%	100%	100%	NA	NA
	Range	78-100%	86-100%	100%	100%	NA	NA
Mexico	Mean	83%	76%	100%	100%	78%	88%
	Range	77-88%	*29-100%	100%	100%	*56-89%	84-100%
U.S.	Mean	72%	83%	100%	100%	87%	89%
	Range	69-81%	*57-100%	100%	100%	78-100%	90-98%
Combined	Mean	83%	82%	100%	100%	83%	89%
	Range	69-100%	29-100%	100%	100%	56-100%	84-100%

Note. DA Pre- and Post = Point-to-point reliability in scoring story grammar from dynamic assessment pretest and posttest. DA Mod = Point-to-point reliability in scoring modifiability rating scale. DA Mod +/- 1 = Reliability in scoring of modifiability rating scale, agreement considered if within 1 point. DA Di = Reliability in dichotomized scoring of modifiability final judgment rating scale. *This low percent agreement is outlying and not reflective of the mean or mode.

indicated that the posttest accounted for 47% of the variance alone, and that the combination of the posttest and the total modifiability variables accounted for 100% of the variance in language ability in the Guatemala students with a Nagelkerke R^2 value of 1.00.

In the second logistic regression model, the posttest score was entered first, followed by the modifiability judgement score, then finally the total modifiability score was entered into the model. The variables were entered in this second model in this particular order to investigate whether the modifiability judgement score was also predictive of language ability in the Guatemalan sample. The results of Model 2 indicated that the combination of the posttest and modifiability judgement variables also accounted for 100% of the variance in language ability. The addition of the modifiability total score to those two variables did not negatively affect the model, with all three variables accounting for all of the variance in language ability (Nagelkerke R^2 = 1.00).

When the gain score variable was combined with posttest scores in the hierarchical logistic regression (Question 1b), it was found to only increase the Nagelkerke R^2 from .48 to .49 and decrease the sensitivity from 75% to 50%. When the duration variable was combined with posttest scores in the hierarchical logistic regression (Question 1b), it was found to only increase the Naglekerke R^2 from .48 to .50 and the sensitivity remained the same at 75%. Because these variables did not appear to significantly predict language ability, they were excluded as potential dynamic assessment predictor variables.

In order to determine the optimal sensitivity and specificity of the Spanish dynamic assessment for the Guatemalan sample (Question 2), ROC analyses, which provided AUC results, were conducted. The AUC provides sensitivity and specificity for each possible cut point of the predictor measure. The predicted probability output from the Model 1-step 2 logistic regression analysis and from the model 2-step 3 logistic regression analysis were used as the predictor measures in the ROC analyses, with language ability as the criterion measure. Results indicated perfect classification accuracy for both Model 1 and Model 2, with an AUC value of 1.00, and 100% sensitivity and 100% specificity.

Because the combination of variables used in the ROC analyses were expressed as logistic regression predicted probability outcomes, the data were also analyzed to identify which clinically interpretable cut-points for the total modifiability score, modifiability final judgement score, and posttest score resulted in the best sensitivity and specificity (Question 3). Results indicated that the best balance of sensitivity and specificity was achieved when students were classified as having a language disorder if they met two or more of the following criteria: (a) a total modifiability score of 19 or lower; (b) a modifiability final judgment score of two or lower; and (c) a posttest score of seven or lower. Using this criterion resulted in sensitivity and specificity values of 100% and 96%, respectively.

Cross-Validating Group 1: Mexico Sample

In order to cross-validate the Guatemalan sample results (Question 4), the same three variables identified in the logistic regression for Guatemala (i.e., Model 2-step 3) were used and applied to the logistic regression for the Mexico sample. This third model (Model 3), with all three significant predictor variables combined, was entered into the hierarchical logistic regression analysis. Results indicated that the three dynamic assessment variables accounted for significant variance in language ability in the Mexico students with a Nagelkerke R^2 value of .74. Following the same procedures used with the Guatemala sample, the classification accuracy of these three variables was also examined using a ROC analysis, which provided AUC and sensitivity and specificity results. The AUC was .96, with 93% sensitivity and 80% specificity, or 87% sensitivity and 100% specificity (depending on whether higher sensitivity or higher specificity is desired).

Additionally, the same clinically interpretable cut-points that were identified in the Guatemala data were applied to the Mexico data with the cutoff for total modifiability at 19 or lower, modifiability final judgment at two or lower, and posttest at seven or lower. This resulted in good classification accuracy, with 80% sensitivity and 93% specificity.

Cross-Validating Group 2: U.S. Sample

Similar to the Mexico sample, the U.S. sample served as an additional cross-validating group to the original Guatemala sample. During the recruitment of the U.S. sample, only two students qualified as having LD according to the qualifications for LD outlined in the method section. Because of this, the U.S. students with LD were not included in the sample and the five students with LD from the Mexico sample were combined with the 20 students with typical language in the U.S. sample to form a total sample size of 25.

The same three predictor variables identified in the logistic regression for the Guatemala sample (i.e., Model 2-step 3) were used with this third sample group. This fourth model (Model 4), with all three significant predictor variables combined (i.e., total modifiability scores, modifiability final judgement scores, and posttest scores) indicated that the dynamic assessment accounted for significant variance in language ability in the U.S. student sample, with a Nagelkerke R^2 value of .69. When the classification accuracy of these three variables combined was also expressed using a ROC analysis, which provided AUC and sensitivity and specificity results, the area under the curve was .91 with 100% sensitivity and 80% specificity.

The same clinically interpretable cut-points identified first with the Guatemala sample and applied to the Mexico sample (i.e., total modifiability at 19 or lower, modifiability final judgment at 2 or lower, and posttest at 7 or lower), were also used with this third sample of U.S. data. When these cut points were applied, sensitivity and specificity values were found to be 80% and 100% respectively.

Guatemala, Mexico, and U.S. Samples Combined

A final set of analyses were used to investigate the classification accuracy of the Spanish dynamic assessment as a whole, across all three sample groups, and to determine how well the Spanish narrative dynamic assessment accounted for language variance across these three groups (Question 5). To answer this question, the same criteria used for the Guatemalan data (Model 2-step 3) were applied in the logistic regression for all of the groups combined. This model (Model 5), indicated that the dynamic assessment accounted for significant variance in language ability across all three samples with a Nagelkerke R^2 value of .69. When the classification accuracy of these three variables combined was expressed using a ROC analysis, results indicated that the AUC was .94, with 100% sensitivity and 78% specificity, or 89% sensitivity and 89% specificity (depending on the extent to which sensitivity or specificity are valued). Lastly, when the same clinically interpretable cut-points used for all three sample groups independently were applied to this final, combined group, good classification accuracy was achieved resulting in 80% sensitivity and 100% specificity. Logistic Regression and ROC analysis results for the combined sample, along with the three other samples, can be found in Table 4.

Discussion

Previous studies have investigated the classification accuracy of narrative dynamic assessments for monolingual English speakers and bilingual Spanish-English speakers, however, at this time, no narrative dynamic assessment designed specifically for monolingual Spanish speakers has been investigated. The purpose of this study was to determine the classification accuracy of a Spanish narrative dynamic assessment of language using a sample of Guatemalan Spanish-speaking students and then to cross-validate those results across two additional samples of Spanish speaking students (i.e., students from Mexico and from the U.S.). In doing so, it was hypothesized that modifiability scores and posttest variables would be the most predictive dynamic assessment variables of language ability, accounting for the greatest amount of variance and achieving the best balance of sensitivity and specificity.

Classification Accuracy of the Spanish Narrative Dynamic Assessment

To determine the classification accuracy of the Spanish dynamic assessment of language, five dynamic assessment measurement variables were initially investigated – total modifiability scores, modifiability final judgement scores, posttest scores, gain scores, and duration of the teaching phase. Previous dynamic assessment research indicated that each of these five measures could be predictive of language ability, but that posttest scores combined with modifiability scores were the most consistently predictive measures (Kramer et al., 2009; Peña et al., 2006; Peña et al., 2014; Petersen et al., 2017; Ukrainetz et al., 2000). Results of the Guatemalan data analysis aligned closely with previous dynamic assessment research, as posttest scores and modifiability measures combined accounted for 100% of variance in language ability, resulting in 100% sensitivity and 100% specificity.

Gain scores and duration scores were not found to meaningfully contribute to the predictive model. However, duration of the teaching phase could be considered an indicator of student language learning ability. For example, a student with a LD, who finds it difficult to learn narrative language skills, would be expected to have a considerably longer teaching phase duration due to the need for more prompting, teaching, and repetition, compared to a student with typical language development who requires less support. A similar case might be made for gain scores. Consistent with the theory of dynamic assessment, one would expect students with typical language development to make larger gains from pre- to posttest as they learn and apply narrative language skills compared to students with LD who struggle to learn and master narrative language skills. It is important to note that out of the five initial dynamic assessment measures investigated, the results of this study suggest that posttest and modifiability measures are the strongest indicators of language ability; however, this does not mean that duration of the teaching phase and gain scores cannot be indicative of language ability, but that they may not be as consistent of classification predictors as posttest and modifiability measures. Gain score and duration measures could still be of significant clinical value to clinicians assessing and formulating treatment to fit the needs of specific students.

Table 4

Logistic Regression and ROC Analyses for Spanish Narrative Dynamic Assessment Predictor Variables for the Guatemala, Mexico, U.S., and Combined Samples

Criterion Measure	Model	Step	Predictor	Beta	Exp(B)	R^2	ΔR^2	χ^2	Wald	Sens.	Spec.	AUC
Guatemala Language Ability	1	1	Posttest	- .43	0.65	.47		8.38**	0.001	1.00	.75	.87
		2	Mod Total	-40.98	0.00	1.00	.53	14.28**	0.001	1.00	1.00	1.00
	2	1	Posttest	-2.49	0.08	.47		8.38**	0.00	1.00	.75	.87
		2	Mod Judge	-26.74	0.00	1.00	.53	14.29**	0.00	1.00	1.00	1.00
		3	Mod Total	.69	1.99	1.00	.00	0.00**	0.00	1.00	1.00	1.00
Mexico Language Ability	3	1	Posttest	.38	1.46	.38		6.00**	1.80	.93	.60	.80
		2	Mod Judge	-.90	.41	.39	.01	.10	1.78	.93	.80	.88
		3	Mod Total	7.13	1243.41	.74	.35	7.74**	2.62	.93/.87	.80/1.00	.96
U.S. Language Ability	4	1	Posttest	.31	1.37	.35		5.97**	2.96	.90	.60	.79
		2	Mod Judge	-.03	0.97	.67	.32	7.11**	.00	1.00	.80	.89
		3	Mod Total	.48	1.61	.69	.02	0.62	.49	1.00	.80	.91
All Kids Combined	5	1	Posttest	.32	1.38	.39		15.52**	5.75*	.95	.67	.83
		2	Mod Judge	3.27	26.25	.65	.26	13.31**	4.84*	.93/.97	.89/.78	.93
		3	Mod Total	-.28	.75	.69	.04	2.47	1.88	1.00/.89	.78/.89	.94

Note. Posttest = dynamic assessment posttest total score. Mod Total = dynamic assessment modifiability total score. Mod Judge = dynamic assessment modifiability final judgment score. AUC = area under the curve. **$p \leq .01$; *$p < .05$. Beta, Wald, and Exp(B) (odds ratio) are from the last step of each model. χ^2 degrees of freedom are equal to the number of predictors in each model.

Cross-Validation Results of the Spanish Dynamic Assessment of Language

Because posttest, total modifiability, and final modifiability judgement variables were found to be most predictive of language ability with the Guatemalan sample, these same variables were applied to the cross-validation Mexico and U.S. samples, and then to all three samples combined. Results indicated that across all cross-validated samples, high percentages of variance in language ability were accounted for, and good classification accuracy was achieved. For example, when all three dynamic assessment variables were applied to the Mexico sample, results indicated that 74% of variance in language ability was accounted for with good classification accuracy (93% sensitivity with 80% specificity or 87% sensitivity and 100% specificity). The same process was applied to the U.S. sample with results indicating 69% of variance in language ability being accounted for while achieving perfect sensitivity and good specificity (100% sensitivity and 80% specificity). When all three samples (i.e., Guatemala, Mexico, and U.S.) were combined, 69% of variance in language ability was accounted for, and overall good classification accuracy achieved, with sensitivity and specificity results of 100% sensitivity with 78% specificity or 89% sensitivity and 89% specificity.

The results found with each individual cross-validation sample and when all three samples were combined is promising. Applying the same three dynamic assessment variables found to be highly predictive in the Guatemala sample to the Mexico and U.S. samples, provided insight into whether or not posttest scores and modifiability measures continued to remain predictive when applied to two additional sets of independent, geographically and culturally diverse Spanish-speaking students. Although, perfect classification accuracy was not achieved with the Mexico, U.S. samples, and combined samples, the sensitivity and specificity values found for both samples remained at or above what is considered to be acceptable (i.e., 80%

sensitivity and 80% specificity; Swets, 1988; Spaulding et al., 2006). The results of the Mexico and U.S. cross-validation samples, as well as the combined sample consisting of all three groups, only add to a wealth of previous research indicating that the combination of dynamic assessment posttest scores and modifiability measures are highly accurate at identifying diverse school-age children with and without LD (Peña et al., 2006; Peña et al., 2014; Petersen et al., 2017).

Clinically Interpretable Cut Point Results

When clinically interpretable cut points, resulting in the best balance of sensitivity and specificity, were found with the Guatemala sample, and then applied to each cross-validating sample, the classification accuracy of the dynamic assessment maintained a high level of accuracy. When these cut-points were applied to the Guatemalan sample, sensitivity remained at 100% with 96% specificity. When applied to the cross-validating samples, 80% sensitivity of and 93% specificity of was achieved with the Mexico sample and 80% sensitivity and 100% specificity with the U.S. sample. Results from the combined sample group revealed 80% sensitivity and 100% specificity. The relative consistency between the sensitivity and specificity values as determined by clinically selected cut-points, when compared to sensitivity and specificity values determined by ROC analysis indicate that cut-points using posttest, total modifiability, and modifiability final judgement scores were clinically useful indicators of language ability when applied to this specific sample of Guatemalan, Mexican, and U.S. Spanish-speaking students. As the ultimate goal is for the Spanish narrative dynamic assessment to be used to clinically identify students with LD, providing evidence of validity of the Spanish dynamic assessment using these clinically appropriate cut points is significant.

Clinical Application of the Dynamic Assessment

The current need for more valid assessments of language in Spanish. To avoid linguistic and cultural biases resulting in the over-identification of CLD students, the U.S. Department of Education mandated that all students be assessed in their native language and that diagnostic materials should be culturally appropriate, avoiding any culturally or racially discriminatory items or materials (Individuals with Disabilities Education Act, 2004) Currently, there is no Spanish narrative dynamic assessment of language designed specifically for monolingual and dominant Spanish-speakers. As the Spanish-speaking student population increases throughout the world and in the United States, an accurate and culturally appropriate assessment of language disorder is needed. The results of this study indicate that a Spanish narrative dynamic assessment of language has the potential to be used clinically to fill this Spanish assessment gap. The high levels of classification accuracy found with each independent, cross-validating sample, and with all participants combined, provides strong preliminary evidence that the Spanish narrative dynamic assessment may be a valid and reliable measure of LD in Spanish dominant students and bilingual Spanish speaking students. The evidence of validity of this assessment was further strengthened as high classification accuracy was maintained when clinically derived cut-points were used to classify students with and without LD. Additionally, the cross-validating nature of this study, with the administration of the dynamic assessment taking place in Guatemala, Mexico, and the U.S., indicates that the Spanish narrative dynamic assessment may yield results that are meaningful and interpretable when used to assess a variety of diverse Spanish-speaking students.

Although, the results of this study are promising, perfect classification accuracy across all sample groups was not achieved. For this reason, and because best practice indicates that a combination of formal and informal assessments be used to successfully identify students with and without LD, it is recommended that the Spanish narrative dynamic assessment be used as only one tool, of many, to help triangulate the presence or absence of LD in CLD and Spanish-speaking populations.

Common clinician concerns related to dynamic assessment. Previously, many clinicians have been concerned with the long administration lengths and the subjective nature of scoring associated with dynamic assessments of language (Deutsch & Reynolds, 2000; Hasson, 2007; Haywood & Tzuriel, 2002). Many of these concerns are well founded. For example, in Peña et al. (2006), the pretest and posttest required language sample transcription and analysis and the teaching phase of the narrative dynamic assessment consisted of two mediated learning experiences of 30 minutes each, after which administrators were asked to complete two separate modifiability scoring protocols. Similarly, Ukrainetz et al. (2000) used a test-teach-retest dynamic assessment format over a period of three weeks with pretest and posttest phases taking 20 minutes each to administer as well as the completion of two separate 30-minute teaching phases. Recognizing that long administration times are out of line with the realities of many school-based SLP workloads, efforts were made in this study to shorten and condense the format of the Spanish narrative dynamic assessment—with pretest and posttest phases taking approximately 3-5 minutes each and a single teaching phase taking approximately 5-10 minutes—while still maintaining high classification accuracy.

In regard to clinician concerns relating to the subjective nature of dynamic assessments scoring, efforts were made to standardize the teaching phase of the Spanish narrative dynamic assessment, with instructions on when and how to use clear and specific prompts to guide student learning detailed on the protocol sheet. Furthermore, to improve clinician confidence when assigning total modifiability and modifiability final judgement scores, the modifiability rating sheet provided specific criterion students were to meet to be assigned a certain score. The high levels of interrater reliability when scoring student modifiability achieved in this study revealed that the more perceived subjective nature of dynamic assessment did not decrease the accuracy and reliability of the Spanish narrative dynamic assessment. These results are consistent with other narrative dynamic assessment of language studies (Kramer et al., 2009; Peña et al., 2006; Peña et al., 2014; Petersen et al., 2017).

The role of dynamic assessment in informing treatment. In addition to the role narrative dynamic assessments of language can play in identifying LI in CLD populations, dynamic assessments of language can also provide valuable insight to clinicians that will later inform their treatment with identified students. One benefit of using dynamic assessment is that it allows clinicians to gage students' language learning abilities, particularly during the teaching phase. The teaching phase allows clinicians to trial a mini treatment session. By doing so, clinicians can better understand how quickly individual students can master targets and what kinds of support and prompting will be required for them to do so. By synthesizing and reflecting on this information, clinicians can identify potential targets for follow-up treatment sessions.

Limitations of the Study

Overall, this study found that a Spanish narrative dynamic assessment was accurate in identifying Guatemalan students with and without LD. Results of the cross-validation analyses further determined that good classification accuracy was maintained, and a large degree of variance in language ability was accounted for, when the Spanish narrative dynamic assessment was administered to two cross-validation samples. However, limitations regarding sample groups and the initial method used to determine LD may have influenced the results of this study.

Participant limitations. First, students in the cross-validating Mexico and U.S. samples were not matched to students in the Guatemalan sample. Student grades in Guatemala ranged from kindergarten to twelfth grade. In comparison, the Mexico and U.S. samples were limited to only elementary school-age students ranging from first to sixth grade. While there is a degree of cross-over in elementary school student ages across the three samples, a large percentage of Guatemalan students were middle to high-school-age students with the average student grade level being eighth grade compared to the average grade level of the Mexico and U.S. students being second and fifth grades respectively. Additionally, while some information regarding students' SES was obtained across all samples, based on the information gathered, it is unclear how closely matched students were in terms of maternal education level and the number of years the students were enrolled in and attended school. Differences in students' ages and socioeconomic status may have influenced their performance on the Spanish narrative dynamic assessment. While posttest scores, total modifiability scores, and modifiability final judgement scores were found to predict the language ability of the Guatemalan sample with perfect accuracy, variations in average sample age and socioeconomic status in the cross-validation samples may have influenced the predictive accuracy of these same dynamic assessment measures with the Mexico and U.S. students. The dynamic assessment variables and the cut-points used for the Guatemalan sample were applied generally to the other cross-validation groups. It is possible that different dynamic assessment variables and different cut-points would have better classified the students in the cross-validation samples. However, while there were variations in the classification accuracy of the dynamic assessment across sample groups, overall good classification accuracy was maintained with each sample group individually and when all three samples were combined.

In addition to limitations regarding sample matching, only a small number of students met the research qualifications set to be considered a student with LD. In the Guatemalan sample, only four out of 28 children were identified as having LD. Similarly, in the Mexico sample, the total sample consisted of 20 students with only five children being students with LD. This small ratio of students with LD to students with typical language had a significant impact on the accuracy of the dynamic assessment in terms of sensitivity and specificity. With such a small number of students with LD per sample, a single outlier score could add to or decrease from the overall sensitivity by 25% in the Guatemalan sample and by 20% in the Mexican and U.S. samples. For this reason, conservative cut off points had to be selected in order to achieve the best balance of sensitivity and specificity. However, it should also be noted that due to the small number of students with LD it is impressive that such high sensitivity and specificity results were found across all sample groups.

Lastly, while two students in the U.S. sample were initially identified as students with LD, these students were removed from the sample due to the small ratio of LD to TD students. For this reason, the five LD students in the Mexico were supplemented into the U.S. sample to improve the LD to TD ratio. Combining the Mexican students with LD with the U.S. TD group directly influenced the cross-validation analysis of the U.S. sample, especially the sensitivity results found using the clinically interpretable cut-points. Because the students with LD in both the Mexico and U.S. samples were the same students, when these clinically-based cut points were used, sensitivity results of both the Mexico and U.S. sample was 80%. Because the Mexican students with LD were used in this third sample group, the U.S. sample was not a true cross-validation sample in terms of students with LD (see Peña et al., 2014 for similar methodology). However, the students with typical language did act as a true cross-validating group with the reported specificity results being true indicators of the classification accuracy of the Spanish narrative dynamic assessment for bilingual English-Spanish-speaking students with typical language.

Variations in the methods used to determine language status. Study limitations also included variations in the method used to initially identify students with LD prior to the administration of the dynamic assessment. Due to variations in available resources and needed adaptations necessary to collaborate with multiple teams when gathering data in all three geographical locations, the measures used to identify LD were not completely consistent from one sample group to the next. For example, in Guatemala, students produced a language sample using the picture book *Frog, Where Are You?* (Mayer, 1969) compared to use of the Narrative Language Measure (NLM; Peterson & Spencer, 2016) to obtain language samples from the Mexico and U.S. students.

Additionally, variations in special education organizations across all three locations required flexibility when seeking confirmation of language status from professionals who had personally associated with the students. For example, the requirement that students be on an IEP to qualify as having a LD could not be met by students in the Mexico and Guatemala samples in which educational documentation differs. In these locations, this qualification was considered to be met if the students' special educators affirmed the student had difficulty acquiring language skills and were receiving formal services for language. Although, identification methods differed moderately from one sample to the next, great care was taken to ensure that the students met multiple critical indicators of LD, even though the exact methods for meeting these indicators varied from sample to sample. With each student identified has having LD, researchers used a systematic process using multiple measures of LD to confidently determine language status and assign students to the LD or TD groups.

Limitations related to outlier interrater reliability results. Outlier interrater reliability results for the NLM and Spanish narrative dynamic assessment modifiability measures directly influenced overall mean interrater reliability results for the Mexico and U.S. samples. For example, an outlying interrater agreement of 29% was found for a single participant in the Mexico sample when rating the dynamic assessment modifiability measure. In another case, when completing interrater reliability for the NLM, two independent raters only agreed point-to-point 57% of the time. These outlier scores may be reflective of variations in training across the three separate examiner teams. For example, examiners in the Mexico team were trained by a collaborating researcher in Mexico, while examiners in the U.S. and Guatemala team were trained by a Spanish-English bilingual, certified SLP from the U.S. However, while a few outlier scores were found, mean interrater reliability in scoring across all three samples and across all measures remained relatively high. Additionally, as two of the three outlier scores found related to the scoring of dynamic assessment modifiability, additional analyses of interrater agreement were performed for modifiability scoring (i.e., dichotomized agreement and agreement when scoring was within one point of each other). These additional analyses resulted in 100% interrater reliability for dynamic assessment scoring of modifiability across all samples.

Future Research

This was an initial investigation into the classification accuracy of a Spanish narrative dynamic assessment. The results of this study indicate that a Spanish narrative dynamic assessment may be used to confidently identify Spanish-speaking students with LD from a variety of cultural backgrounds. However, due to the limitations of the study, and because the study sample consisted of only a relatively small number of students living in three geographical regions, the results of this study cannot be assumed to apply to all Spanish-speaking students. Future research is needed to investigate the classification accuracy of a Spanish narrative dynamic assessment with a larger sample size of Spanish speakers of various ages, cultural backgrounds, and linguistic abilities. It would be of particular interest to discover whether posttest and modifiability scores continue to be most predictive of language ability across this wider and more representative sample, or if other dynamic assessment variables add to the accuracy of the dynamic assessment. Additionally, as the cut-points used in this study were applied to a wide age range (i.e., kindergarten through twelfth grade), it would be of interest to determine if high sensitivity and specificity could be maintained across a larger sample of kindergarten-twelfth grade students using these same cut-points.

Conclusion

The purpose of this study was to investigate the classification accuracy of a Spanish narrative dynamic assessment with a single sample of Guatemalan students with and without LD, and to then cross-validate those results over two additional samples of Spanish-speaking students. In doing so, dynamic assessment posttest scores when combined with total modifiability and modifiability final judgement scores resulted in 100% sensitivity and specificity for the Guatemalan sample. When these same dynamic assessment predictor variables were applied to the two cross-validation samples and the Guatemalan and cross-validation samples combined, high classification accuracy was maintained. These results indicate that the Spanish narrative dynamic assessment of language is able to accurately differentiate between Spanish-speaking Guatemalan, Mexican, and U.S. students with and without LD. Furthermore, the results of this study provide strong preliminary evidence that a Spanish narrative dynamic assessment may be a valid assessment when identifying monolingual and dominant Spanish-speaking students in the U.S. and Spanish-speaking countries.

References

Archibald, L. M. D., & Gathercole, S. E. (2006). Nonword repetition: A comparison of tests. *Journal of Speech, Language, and Hearing Research, 49*, 970-983.

Barragan, B., Castilla-Earls, A., Martinez-Nieto, L., Restrepo, M., & Gray, S. (2018). Performance of low-income dual language learners attending English-only schools on the clinical evaluation of language fundamentals (4th ed., Spanish). *Language, Speech, and Hearing Services in Schools*, *49*, 292-305. https://doi.org/10.1044/2017_LSHSS-17-0013

Bauman, K. (2017). *School enrollment of the Hispanic population: Two decades of growth.* Retrieved from https://www.census.gov/newsroom/blogs/random-samplings/2017/08/school_enrollmentof.html

Boudreau, D. (2008). Narrative abilities: Advances in research and implications for clinical practice. *Top Language Disorders, 28,* 99-114. doi: 10.1097/01.TLD.0000318932.08807.da

Cervantes Institute. (2016). *El Espanol: Una lengua viva*. Retrieved from https://www.cervantes.es/imagenes/File/prensa/EspanolLenguaViva16.pdf

Colozzo P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54*, 1609-1627. doi: 10.1044/1092-4388(2011/10-0247)

Deutsch, R., & Reynolds, Y. (2000) The use of dynamic assessment by educational psychologists in the UK. *Educational Psychology in Practice*, *16*, 311-331.

Dziuban, C. D., & Vickery, K. V. (1973). Criterion-referenced measurements: Some recent developments. *Educational Leadership*, *30*, 483-486.

Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers*. Baltimore, MD: University Park Press.

Gathercole, S. E., & Baddeley, A. D. (1996). *The Children's Test of Nonword Repetition*. London, England: Psychological Corporation.

Hasson, N. (2007). The case for dynamic assessment in speech and language therapy. *Child Language Teaching and Therapy*, *23*, 9-25.

Haywood, C., & Tzuriel, D. (2002). Applications and challenges in dynamic assessment. *Peabody Journal of Education, 77*, 40–63.

Individuals with Disabilities Education Act, 20 U.S. C. § 1414 (2004).

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools, 43*, 81-96. doi: 10.1044/0161-1461(2011/10-0095)

Kramer, K., Mallett, P., Schneider, P., & Hayward, D. (2009). Dynamic assessment of narratives with grade 3 children in a first nations community. *Canadian Journal of Speech-Language Pathology and Audiology/Revue Canadienne d 'Orthophonie Et d'Audiologie, 33*, 119-128.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools, 34,* 44-55.

Mayer, M. (1969). *Frog, where are you?* New York, NY: Dial Press.

Miller, J. Gillam, R. B., & Peña, E. D. (2001). *Dynamic Assessment Instrument*. Austin, TX: Pro-Ed.

Miller, J., & Iglesias, A. (2012). Systematic Analysis of Language Transcripts (SALT), Research Version 2012 [Computer Software]. Middleton, WI: SALT Software, LLC.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*, 2208-2220. doi: 10.1044/2014_JSLHR-L-13-0151

Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*, 1037-1057. doi: 10.1044/1092-4388(2006/074)

Peña, E. D., Quinn, R., & Iglesias, E. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *Journal of Special Education, 26,* 269-280.

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*, 983-998. doi: 10.1044/2016_JSLHR-L-15-0426

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48,* 3-21. doi: 10.1044/cds21.1.5

Petersen, D. B., & Spencer, T. D. (2014). *Predictive Early Assessment of Reading and Language* (PEARL). Laramie WY: Language Dynamics Group.

Petersen, D. B., & Spencer, T. D. (2016). *The CUBED: Narrative Language Measures* (NLM). Laramie WY: Language Dynamics Group.

Retrieved from http://www.languagedynamicsgroup.com/.

Popham, J. W., & Husek, T. R. (1969). Implications of criterion referenced measurement. *Journal of Educational Measurement 6*, 1-9.

Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10,* 382-393.

Semel, E., Wiig, E. H., Secord, W. A. (2006). *Clinical Evaluation of Language Fundamentals-Fourth Edition, Spanish* (CELF-4 Spanish). Bloomington, MN: NCS Pearson.

Spaulding, T. J., Plante, E., Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61-72.

Squires, K. E., Lugo-Neris, M., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language & Communication Disorders, 49*, 60-74. doi: 10.1111/1460-6984.12044

Swets J. A. (1988). Measuring the accuracy of dynamic systems. *Science, 240*, 1285-1293.

Tomblin, J. B., Records, N. L, Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research*, *40*, 1245–1260.

Ukrainetz, T. A., Stacey, H., Walsh, C., & Coyle, C. (2000). A preliminary investigation of dynamic assessment with Native American kindergarteners. *Language, Speech, and Hearing Services in Schools, 31*, 142-154.

U.S. Census Bureau (2011). *The Hispanic population: 2010*. Retrieved from https://www.census.gov/prod/cen2010/briefs/c2010br-04.pdf

U.S. Census Bureau (2018). *Annual estimates of the resident population by sex, age, race, and Hispanic origin for the United States and states: April 1, 2010 to July 1, 2017*. Retrieved from https://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?src=bkmk#.

Vespa, J., Armstrong, D. M., & Medina, L. (2018). Demographic turning points for the United States: Population projections for 2020 to 2060. In *U.S. Census Bureau.* Retrieved from https://www.census.gov/content/dam/Census/library/publications/2018/demo/P25_1144.pdf

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental processes*. Cambridge, MA: Harvard University Press.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1993). *Preschool Language Scale-3: Spanish Edition*. San Antonio, TX: Psychological.

APPENDIX A

Modifiability Rating Scale

Modificabilidad					
Puntos	**4**	**3**	**2**	**1**	**0**
Respuestas a indicaciones	El alumno responde a las indicaciones la mayer parte del tiempo. Los avisos son principalmente de Nivel 1.		El alumno responde a las indicaciones parte del tiempo. Los apuntes son principalmente de Nivel 2.		El alumno responde a las indicaciones sóla rara vez. Todos los apuntes son de Nivel 2.
Gado de transferencia	1-2 objetivos se transfieren a través de ciclos. Todos los elementos de la historia grammatical se incluyen en el ciclo final.		Por lo menos, un objective se transfiere a veces a través de ciclos. Muchos elementos de la historia grammatical se incluyen en el ciclo final.		La transferencia de objetivos a través de ciclos es rara. Poco elementos de la historia grammatical se incluyen en el ciclo final.
Atención a enseñanza	Atento y centrado. El alumno no require redirección verbal.		En la tarea a veces. El alumno require algunas redirecciones verbales y puede ser reenfocado.		Distraido y dificil. Se necesita una redirección significative.
Facilidad enseñando	El alumno require esfuerzo minimo del examinador para ver cambio. El esfuerzo del examinador disminuye a través de ciclos.		Un poco esfuerzo del examinador es requerido para ver cambio. El esfuerzo del examinador disminuye un poco a través de ciclos.		Mucho esfuerzo del examinador es requerido para ver cambio. El esfuerzo del examinador disminuye muy poco a través de ciclos.
Nivel de frustración	Muy poco o no frustración está persistente y mantience la atención fácilmente a las tareas.		Poca frustración está indicado. Está vacilante y timido. Tal vez requiere consuelo.		Mucha frustracion está exhibida. Está afigido y requiere mucho consuelo.
Interrupciones	Muy poco o no comportamiento verbal que interrumpe la intervención.		Poco comportamiento verbal que interrumpe la intervención.		Mucho comportamiento verbal que interrumpe la intervención.
Crítica final del examinador					
Puntos	**4**	**3**	**2**	**1**	**0**
Crítica final del examinador de la capacidad del estudiante para aprender lenguaje					

APPENDIX B

Parental Consent Form

Petersen Dynamic Assessment: Spanish

Permiso de los Padres para un Menor

Introducción

Mi nombre es Douglas Petersen. Soy profesor de la Universidad de Brigham Young. Estoy haciendo una investigación sobre lenguaje. Quiero invitar a su hijo a participar en la investigación porque está en la escuela primaria y necesitamos ver si nuestra nueva evaluación de lenguaje puede medir el lenguaje con precisión para todos los estudiantes en edad escolar.

Procedimientos

Si permite que su hijo participe en esta investigación, ocurrirá lo siguiente:

-Se le pedirá a su hijo que cuenta cuentos.
-Su hijo(a) puede contar estos cuentos en el aula, fuera del aula en la escuela, o en casa.
-Su hijo(a) va a aprender contar cuentos.
-Su hijo(a) estará administrado un examen de lenguaje estandarizado.
-Pasaremos aproximadamente 45 minutos a poco más de una hora con su hijo(a).
- Solicitaremos información sobre su hijo de su escuela, incluidos su etnia/raza, almuerzo gratis/reducido, habilidad de hablar inglés, y si recibe educación especial.

Riesgos

Existe riesgo de pérdida de privacidad, que el investigador reducirá al no utilizar nombres reales u otros identificadores en los manuscritos. El investigador también mantendrá todos los datos en un archivador bloqueado en un lugar seguro. Solo el investigador tendrá acceso a los datos. Al final del estudio, los datos se compartirán con usted y con el maestro de su hijo.

Puede haber alguna molestia por su hijo(a) contar cuentos o recibir examines de lenguaje. Su hijo(a) puede detener todo el proceso en cualquier momento sin afectar su posición en la escuela o las calificaciones en clase.

Confidencialidad

Los datos de la investigación se mantendrán en un lugar seguro (o protegido con contraseña y encriptado) y solo el investigador tendrá acceso a los datos. Al finalizar el estudio, se eliminará toda la información de identificación y los datos se guardarán en un gabinete u oficina cerrada.

Beneficios

Las evaluaciones de lenguaje pueden ayudarlo a usted o al maestro a identificar la dificultad con el lenguaje o la comprensión de lectura.

Compensación

No habrá compensación por la participación en este proyecto.

Preguntas sobre la investigación

Petersen Dynamic Assessment: Spanish
Por favor dirija cualquier pregunta sobre la investigación a **Douglas Petersen: dpeter39@byu.edu, 435-213- 5262.**

Las preguntas sobre los derechos de su hijo(a) como participante en la investigación o para enviar comentarios o quejas sobre la investigación deben dirigirse al Administrador del IRB, Universidad de Brigham Young, A-285 ASB, Provo, UT 84602. Llame al (801) 422-1461 o envíe correos electrónicos a irb@byu.edu.

Se le ha entregado una copia de este formulario de consentimiento para conservar.

Participación

La participación en esta investigación es voluntaria. Usted es libre de negarse a que su hijo participe en esta investigación. Puede retirar la participación de su hijo en cualquier momento sin afectar el grado / posición de su hijo en la escuela, el tratamiento o los beneficios.

El Nombre del Niño: ______________________________

Nombre de Padres:_______________ Firma:__________________ Fecha:____________

Si elige que su hijo participe en esta investigación, indique si le preocupa el desarrollo del lenguaje de su hijo/a:

x Me preocupa el dsarrollo del lenguaje de mi hijo/a en *inglés*.

x NO me preocupa el dsarrollo del lenguaje de mi hijo/a en *inglés*.

x Me preocupa el dsarrollo del lenguaje de mi hijo/a en *español*.

x NO me preocupa el dsarrollo del lenguaje de mi hijo/a en *español*.

APPENDIX C

Annotated Bibliography

Boudreau, D. (2008). Narrative abilities: Advances in research and implications for clinical practice. *Topics Language Disorders, 28,* 99-114. doi: 10.1097/01.TLD.0000318932.08807.da

Objective: Boudreau (2008) organized and summarized the latest research involving narrative language and school age children in an effort to help instruct clinical practice. *Results:* After surveying the available literature, Boudreau determined that early child narrative abilities are predictive of later academic and social success in all children including children with language impairment. Trends in narrative research indicate that children with high levels of narrative language receive higher scores on academic measures, including math scores, than children with low narrative scores. The research also suggests that low scores on narrative measures in kindergarteners are one of the best predictors of the future need for academic remediation. The author also found that challenges with narrative language are a "persistent characteristic of language impairment". The narratives of children with language impairment consistently show shorter utterance lengths, less grammatical complexity, fewer story grammar elements, fewer story episodes, and overall reduced quality when compared to the narratives of typically developing peers and that the narrative capabilities of children with LI generally persist below the levels of typical peers through their school-age years. *Relevance to current work:* This study supports the assertion that narrative skills are vital language skills. Less complex narrative language skills in kindergarten are indicative of language disorder and are predictive of later academic success.

Colozzo P., Gillam, R. B., Wood, M., Schnell, R. D., & Johnston, J. R. (2011). Content and form in the narratives of children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 54*, 1609-1627. doi: 10.1044/1092-4388(2011/10-0247)

Objective: The goal of this study was to analyze the narratives of typically developing children and the narratives of children with language impairment in terms of grammatical accuracy and quality of content. *Methods:* Two sets of participants were used in the study. One group of participants was drawn from schools in British Columbia, Canada. Twenty-six children (13 with LI, and 13 TD) were selected from second and fourth grade to participate in the study. All of the participants in the first group were considered monolingual English speakers. Participants were determined to have LI if they were receiving oral language intervention and scored at or below a 7 on two of three subtests of the Clinical Evaluation of Language Fundamentals (CELF-3; Wigg, Semel, & Secord, 2000). Each participant completed two narrative production tasks taken from the Test of Narrative Language (TNL; Gillam, Pearson, 2004) in which they were asked to generate a story from a picture sequence or single picture. The narratives were audio recorded and then transcribed. Each narrative was given a narrative content score and a narrative form score. Interrater reliability for the scoring of the picture-sequencing task was 96% and 87% for the single picture task. The second group of participants was taken from the data of 40 children (20 LI, 20 TD) living in Texas or Kansas who had previously participated in a clinical trial. This second set of participant data was examined for the purposes of replication. The Texas/Kansas participants had previously been assessed using the TNL and their scores from the same subsections were analyzed using the same format of the British Columbia study. Interrater reliability for TNL scoring of this group was 94%. *Results*: Results from the British Columbia participant group indicated that the LI group scored much lower on the narrative form and content scores than compared to the TD group. Additionally, participants with LI tended to have stronger content in their stories than grammar or vice versa. Rarely were the participants' grammatical scores and content scores matched in complexity. In the replication study, the results also indicated that the children with LI performed significantly lower in scores of narrative content and narrative form compared to their typical peers. Lastly, the Texas/Kansas LI participants also showed the same scoring pattern as the British Columbia participants, scoring significantly higher in one score, either the content or form score, but markedly lower in the other. *Relevance to current work:* This study supports the assertion that children with LI have less complex narrative skills, in both form and content, when compared to the narrative skills of children with typically developing language.

Fey, M. E., Catts, H. W., Proctor-Williams, K., Tomblin, J. B., & Zhang, X. (2004). Oral and written story composition skills of children with language impairment. *Journal of Speech, Language, and Hearing Research, 47*, 1301-1318.

Objective: The purpose of this study was to follow four groups of school age children (children with persistent language impairment, children with indeterminate language impairment, children with low nonverbal IQ, and typically developing children) from kindergarten to fourth grade while monitoring their development of narrative language. Research questions included: (a) How do the four groups differ in the gains they make in narrative development over time? (b) What factors are involved in the gains these groups make? And, (c) Does the persistence of LI from kindergarten to second grade predict lower levels of narrative development levels in fourth grade differ. *Method:* Participants of the study were 538 monolingual English kindergarten children from Caucasian, African American, Hispanic, and Native American ethnic backgrounds. 262 of the participants displayed typical language skills, 111 children had language impairment with no cognitive impairment, and 90 had low nonverbal IQs. The narrative retell elicitation procedure was used with each child at three separate times – during kindergarten, second grade, and fourth grade. Each child was given the opportunity to produce two narratives, one written and one oral, using two different picture sets. Prior to producing their narrative, the examiner produced a narrative model using a separate picture set. After modeling, the examiner assisted the child in studying the picture cards and verbally labeling the different story elements portrayed in each picture. After this brief teaching phase, the child was then asked to tell or write the story from start to finish. The narratives were scored and analyzed using six principle measures: (a) The number of different words in the narrative, (b) total number of C-units, (c) the mean length of C-units in words, (d) average number of clauses per C-unit, (e) the percentage of grammatical C-units, and (f) a narrative quality score given by the examiner. Interrater reliability in assigning the narrative quality score was 83%. *Results:* Results indicated that in second and fourth-grade, females out-performed males in most measures. However, while females scored higher in narrative measures in the fourth-grade, the gap between male scores and female scores was narrowing. In both second and fourth grade, students exhibited stronger oral narratives than written narratives. However, gains in written narratives between second and fourth grade were higher than gains in oral narratives. Overall, students with language impairment demonstrated shorter, less complex C-units, fewer different words, and a higher number of grammatical errors than typical language students over time. *Relevance to current work:* This study supports the assertion that the narratives of children with language disorder are less complex than the narratives of students with typical language. Additionally, this study suggests that younger children produce stronger oral narratives than written narratives.

Fiestas, C. E., & Peña, E. D. (2004). Narrative discourse in bilingual children: Language and task effects. *Language, Speech, and Hearing Services in Schools*, *35*, 155-168.

Objective: This study was design to compare and contrast the narratives of bilingual children in both English and Spanish to discover if the narrative skills of the children were more advanced in one language and if cultural influences lead to variations of narratives in one language when compared to the other. Secondly, the study aimed to analyze two types of "narrative elicitation stimuli" to see which stimuli elicited a more complex and complete narrative from the bilingual children. *Methods:* The participants of the study were twelve bilingual (Spanish and English speaking) school-age children. The children were between the ages of 4;0 and 6;11. The twelve bilingual children were selected due to their comparable fluency in both English and Spanish. Four total narratives were elicited and recorded. Two narratives were elicited using two different wordless picture books, one in English and one in Spanish. The two other narratives were elicited using a "visually rich" single picture stimulus with prompts, again one in English and one in Spanish. Each narrative was recorded, transcribed, and then coded for story grammar and segmented into C-units. However, the narratives taken using the picture stimulus were not coded for story grammar because the picture stimulus often did not elicit a true narrative. Story grammar elements were coded using a 7-point rating scale. Each child received one point for each story grammar element they included. The inclusion of story grammar elements was rated by two separate raters. Interrater reliability between the two raters was 91%. A third scorer rated the remaining 9% of disagreements resulting in 100% interrater reliability. C-units were also scored by three different raters resulting in an interrater reliability of 94%. *Results*: The children's use of story grammar elements in both the book and picture tasks in both languages were analyzed to measure story complexity. Overall, the results of the study indicated that the Spanish and English wordless picture book narratives were equally complex. However, in the Spanish book task, the children included an initiating event and attempt more often than in English. In the English book task, the children included a consequence more often than in the Spanish book task. In addition, analysis of the wordless picture book narratives indicated that the children used more Spanish-influenced English utterances than English-influenced Spanish utterances. These differences may be due to differences in narrative style across cultures or may have been due to the different exposure the children had to storytelling in English and Spanish. Lastly, the study determined that the number of words, or productivity, of the Spanish narratives when compared to the English narratives showed little difference between the two languages. *Relevance to current work:* This study shows that if a student is a balanced bilingual the complexity of their narratives are similar in both English and Spanish. Additionally, this study supports the assertion that there are narrative differences across languages due to cultural influences and differences in narrative styles.

Kramer, K., Mallett, P., Schneider, P., & Hayward, D. (2009). Dynamic assessment of narratives with grade 3 children in a first nations community. *Canadian Journal of Speech-Language Pathology and Audiology/Revue Canadienne d 'Orthophonie Et d'Audiologie, 33*, 119-128.

Objective: Kramer, Mallett, Schneider, & Hayward (2009) investigated the classification accuracy of the same Dynamic Assessment Instrument (DAI; Miller et al., 2001) narrative dynamic assessment of language examined by Peña et al. (2006) and Peña et al. (2014) when administered to 17 (12 students with typical language development, 5 students with disordered language development) 3rd grade First Nations students in Alberta, Canada. *Method:* Five of the 17 children who participated in the study were classified as LI while 12 were classified as TD. The children were classified as LI based on input from the special education teacher, 3rd grade teachers, and the school's principal. These professionals provided their input on each of the child's language status using previous language assessments, classroom performance, and their overall observations.

The administration of the dynamic assessment matched the administration procedures as detailed in Peña, 2001 and Peña et al. (2006) and Peña et al. (2014). The same wordless picture books were used as well as the same scoring Likert scales to measure each student's modifiability, responsiveness, and narrative productions as the DAI. The administration of the full DAI took place over a four-day period. Narrative transcripts were each scored by two examiners. The final modifiability, pretest, and posttest scores were assigned following scoring consensus between the two main examiners. However interrater reliability on modifiability scoring and pre- and posttest scores were not reported. *Results:* Results of the study indicated that the dynamic assessment had a high classification accuracy in identifying language impairment among the 3rd grade students. While both the normal language learning students and the atypical language-learning students had similar pretest scores, normal language learning students made greater improvements in targeted and non-targeted narrative elements. The combination of modifiability and gain scores most accurately classified students, yielding 100% sensitivity and 92% specificity. Gain scores alone resulted in the same degree of sensitivity and specificity. Modifiability and responsiveness scores alone resulted in 100% sensitivity but only 75% specificity. *Relevance to current work:* This study shows that dynamic assessments are useful tools in determining language difference from disorder in ethnically diverse students. Because there was a weak process for pre-determining if students had a language impairment, this study shows the importance of having a valid process to select participants with language impairment prior to the administration of the dynamic assessment. The likelihood of a child having language impairment is best decided by triangulating using various measures. If a strong process for determining language ability before assigning participants to the language impairment group or the typical language group is absent, sensitivity and specificity measures may not be accurate. While the classification accuracy of this study is questionable, the study did report high sensitivity and specificity. Modifiability and gain scores together resulted in the highest classification accuracy.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*, 2208-2220. doi: 10.1044/2014_JSLHR-L-13-0151

Objective: The purpose of the study was to determine if an English narrative dynamic assessment would accurately identify LI in bilingual English speakers. *Method:* Participants were 54 bilingual, kindergarten children. Eighteen kindergarteners were determined to have LI and the remaining 36 were typical students. Eighteen of the typical students were matched with the LI group in age, sex, language experience, and IQ. The remaining 18 acted as a comparison group and were matched to the LI group for age and language experience. The dynamic assessment consisted of three total sessions following a test, teach, retest format. The first session included a pretest narrative retelling using a wordless picture book and the first modified learning experience during which the student and clinician reviewed and added missing story elements to the child's pretest story. The second session contained the second modified learning experience during which the clinician and student created a narrative together. The third session included the posttest using another wordless picture book parallel in structure to the wordless picture book used in the pretest phase. The two Mediated Learning Experience (MLE) sessions were 30 minutes each and were conducted entirely in English across two separate days. In the MLE sessions, students practiced producing complete narrative retell episodes. After each MLE, the clinician rated the child's responsivity from 1-5 in 12 areas with a 1 representing little support from the examiner and a 5 representing maximum support from the examiner. The stories themselves were scored using a 5-point scale based on the number of story components included and the quality and complexity of story ideas and language used. Each story was also rated in terms of episode structure using a 7-point scale. Interrater agreement for the scoring of narrative elements the in the pretest and posttest stories was 87%. However, no interrater reliability for modifiability scoring was reported. *Results:* The results of the study showed that when three modifiability variables, three posttest variables, and one language sample variable were combined, the dynamic assessment had a high classification accuracy. Sensitivity and specificity percentages were calculated for the treatment LI group and treatment TD group, for the treatment LI group and a TD matched compare group, and for the treatment LI group and a TD, non-matched group. Across these three different analyses, sensitivity ranged from 89% to 100% and specificity was 89%. *Relevance to current work:* The results of this study support the assertion that English dynamic assessments are accurate in identifying language impairment in bilingual, Spanish-English speakers. Additionally, a combination of modifiability, posttest, and language sample Systematic Analysis of Language Transcripts (SALT; Miller & Iglesias, 2012) scores yielded the highest classification accuracy. It’s also important to note interrater reliability scoring of modifiability was not reported.

Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49*, 1037-1057. doi: 10.1044/1092-4388(2006/074)

Objective: The purpose of this study was two-fold; first, to determine if two wordless picture books were parallel in structure and could both be used as reliable narrative assessment tools and second, to determine the classification accuracy of a narrative language dynamic assessment for LI for English speaking students. The researchers were particularly interested in determining which combination of dynamic assessment scores (pretest, posttest, and/or modifiability) had the highest classification accuracy. *Method:* To address the first question, researchers collected two oral narratives, one for each book, from 58 second-grade students. Results showed similar narrative scores for each book within participants, indicating that the two wordless picture books were parallel in structure. To address the second research question, 71 first and second graders were administered an English dynamic assessment measuring narrative ability. Students came from Latino-American, African American, and European American ethnic backgrounds and were all English speakers. The participants were divided into a control group, a typically developing group, and a language impairment group. Participants were classified as having LI it they met three criterion: parental concern of their child’s language, a score at least 1.25 SDs below the mean on the Test of Language Development-Primary (TOLD-P-3; Newcomer, Hammill, 1997), and previous diagnosis of language disorder by a certified SLP. The dynamic assessment consisted of a pretest phase in which students produced an oral narrative using one of the stories from experiment one. Following the pretest phase the typically developing and language impairment groups participated in two, 30-minute, mediated learning sessions in which narrative components were taught. Two 5-point rating scales were used to score a child's overall modifiability following the MLE sessions. The first scale rated examiner effort while the other rated child responsivity. Interrater reliability for modifiability scoring was not reported in this study. However, Peña, Resendiz, & Gillam (2007) later examined interrater reliability of the modifiability measures. The posttest phase consisted of students producing an oral narrative using the second wordless picture book. *Results:* Results indicated that both the students with LI and the students with typical language develop made gains from pre- to posttest while the control group scores showed little change from pre- to posttest. Modifiability ratings based on examiner effort and child responsivity had the highest sensitivity (93%) and specificity (96%) when compared to the classification accuracy of the pretest or posttest scores. However, when posttest scores and modifiability scores were combined, classification accuracy increased to 100%. *Relevance to current work:* This study shows that the narrative dynamic administered had a high classification accuracy in determining language disorder in diverse English-speaking students. Additionally, the study suggests that sensitivity and specificity are highest when modifiability scores and posttest scores are combined.

Peña, E. D., Resendiz, M., & Gillam, R. B. (2007). The role of clinical judgments of modifiability in the diagnosis of language impairment. *Advances in Speech-Language Pathology, 9*, 332-345. doi: 10.1080/14417040701413738

Objective: The purpose of this study was to examine what aspects of child modifiability, following two narrative dynamic assessment mediated learning experiences, were most predictive of language impairment. In addition, the relationship between modifiability ratings and gain scores were examined, as well as what aspects of clinician to child interactions affected modifiability ratings. *Method:* The participants of this study were 40 children from the Peña et al. (2006) study. These students ranged in age from 6;6 to 8;5 years old. Fifteen of the students were identified as having language impairment while 25 were identified as typically developing. The racial and ethnic profile of the participants included Latino-American, African American, and European American students, and two students whose ethnicity was unknown. The participants with LI were verified using three criteria: (a) They were classified as students with LI by a certified speech-language pathologist; (b) They had a parent or teacher concerned about their language abilities; and (c) They scored 1.25 or more standard deviations below the mean on the TOLD-P:3. The participants participated in two, 30-minute mediated learning sessions. The first session focused on teaching story components and episode structure using the story the children were assessed with during the pretest. The second session used a new story to teach story components and episode structure. After each session, the clinician used the *Mediated Learning Observation* tool used in Peña et al. (2006) to assess child modifiability. This modifiability tool was divided into four headings: affect, cognitive arousal, cognitive elaboration, and behavior. Under the affect heading each child was rated in terms of anxiety, motivation, and non-verbal persistence. Under the cognitive arousal heading students were rating in terms of task orientation, meta-cognition, and non-verbal self-reward. Under the cognitive elaboration heading, students were rated for problem-solving, verbal mediation, and flexibility. Under the behavior heading, students were rating in terms of responsiveness to feedback, attention, and compliance. Examiners rated the child on each subheading using a 5-point scale. Usually, each of the two examiners provided intervention to the student. However, 11 of the students received both intervention sessions from the same examiner. Each rater scored the student's modifiability after each session. Combination of interrater and intrarater reliability was 99% for scoring modifiability. However, the extent that the raters were independent of each other was unclear as they may have used the same protocol form to rate the same student at separate times. *Results:* The results of this study indicated that the typically developing group and the language impairment group differed the most on the cognitive arousal and elaboration modifiability scores. Using the modifiability ratings for cognitive arousal and elaboration resulted in 87% sensitivity and 92% specificity. When only the metacognition and flexibility subheading scores were used, sensitivity increased to 93% and specificity remained at 92%. *Relevance to current work:* The results of this study are strong evidence that modifiability scores can have high sensitivity and specificity. Additionally, the results of the study indicate that a child's personal awareness of their errors and a child's ability to adapt and use new strategies during instruction are highly indicative of the presence or absence of language impairment. Lastly, the metacognition and flexibility sub-ratings may be beneficial to incorporate into future modifiability scoring protocols.

Peña, E. D., Quinn, R., & Iglesias, E. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *Journal of Special Education, 26,* 269-280.

Objective: In a seminal study, Peña and Iglesias (1992) compared the efficacy of dynamic assessment measures of language against standardized language assessment measures in identifying language disorder in culturally diverse populations. In addition, the study also examined the mismatch between the common linguistic tasks found in norm-referenced standardized tests and the linguistic tasks common in Latino American and African American home cultures. *Method:* The participants of the study were 50 African American and Puerto Rican American students from three Head Start preschool classes in Northern Philadelphia. Two standardized, norm-referenced instruments were used: the Expressive One Word Picture Vocabulary Test (EOWPVT; Gardner, 1979) and the Comprehension subtest of the Stanford-Binet Intelligence Scale (CSSB; Thorndike, Hagan & Sattler, 1986). The EOWPVT is a single-word labeling task while the CSSB is a more descriptive test. Researchers hypothesized that the CSSB would be more in line with linguistic tasks more common to the students' home cultures while the EOWPVT tasks would be more foreign the students' cultural and home experiences. Children who scored low on the EOWPVT, a single-word, labeling test, received mediation training. Mediation training consisted of two, 20-minute sessions and focused on improving the labeling abilities of the students. Following each session, the clinicians rated each student in terms of child responsiveness, examiner effort, and transfer to achieve an overall modifiability rating. After the mediation sessions, each student was reassessed using the EOWPVT. *Results:* The results of the study were two-fold. First, following statistical analysis it was determined that both typically developing students and students with language disabilities scored equally low on the EOWPVT during pretesting but that students with language disorders scored significantly lower on the CSSB. Second, the classification accuracy of the dynamic assessment used in the study was determined. Data analysis showed the dynamic assessment was effective in classifying 92% of the language-disordered cases. Lastly, typically developing students had both higher gain and modifiability scores than the students with language impairment. These results indicate that dynamic assessment measures are effective in determining language impairment in culturally diverse students and that initial pre-test standardized measures of assessment are ineffective in discriminating between typically developing and disordered student populations from diverse backgrounds. *Relevance to current work:* This study supports the assertion that dynamic assessments of language are less biased assessment methods than static measures. Static, norm-referenced measures are less effective in discriminating disorder from disability, as many of the test items on these assessments are culturally foreign to diverse students. Additionally, low modifiability scores are indicative of language impairment.

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*, 983-998. doi: 10.1044/2016_JSLHR-L-15-0426

Objective: This study was designed to determine the accuracy of an English narrative dynamic assessment in classifying LI in Spanish-English-speaking, Kindergarten through grade students. The study utilized a condensed narrative dynamic assessment with real time scoring to determine if LI could be identified in a shorter amount of time than traditional dynamic assessment measurements while maintaining adequate classification accuracy. *Method:* The study included 42 Hispanic children (10 with LI and 32 without LI) from an urban location in the mountain west. Each participant was bilingual in both Spanish and English. Their level of proficiency both Spanish and English was determined by analyzing a language sample of each child's speech in terms of mean length of utterance, total number of words, and number of different words. The dynamic assessment consisted of two 25-minute test-teach-retest sessions. Within each session the student participated in a pretest narrative retelling, a teaching phase, and a posttest narrative retelling. The pretest and posttest assessments, and a modifiability rating scale were scored in real time by the examiner. The child's narratives in the pretest and posttest were scored in three main areas: a) inclusion of nine story grammar elements, b) use of coordinating and subordinating conjunctions (then, because, when, and after) and, c) episodic structure. The teaching phase targeted each of the three areas scored in the pre- and posttests. Following both of the teaching phases, the examiner scored the child in terms of their teachability using a modifiability rating scale. *Results:* The results of the study found that the narrative dynamic assessment had high classification accuracy. Modifiability ratings from both dynamic assessment sessions yielded 100% specificity and 100% sensitivity. Additionally, a modifiability rating from one 25-minute session when combined with the duration score yielded 100% sensitivity and 91% specificity. Gain scores were found to be an ineffective method of classification. *Relevance to current work:* The results of this study support the assertion that English dynamic assessments are accurate in identifying language impairment in bilingual, Spanish-English speakers. Results also indicated, that it may be possible to further shorten the dynamic assessment teaching phase to make it even more clinically useful. Lastly, modifiability ratings had a high interrater reliability and were more predictive of language disorder than other methods of scoring.

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities*, 48, 3–21. doi: 10.1044/cds21.1.5

Objective: The purpose of this study was to examine if a dynamic assessment of reading administered to kindergarten bilingual Latino children was predictive of reading difficulty at the end of first grade. A secondary purpose of the study was to compare how accurate the dynamic assessment measure was in identifying participants at risk for literacy problems with the classification accuracy of a more traditional, static measure of reading ability. *Method:* The participants of the study were 63 bilingual Latino-American children in kindergarten. Each participant had previously been identified as at risk for language impairment prior to the study. The students had varying English language abilities but were all considered bilingual and had lived in the US for at least one year. The Bilingual English-Spanish Assessment (BESA; Peña, Gutierrez-Clellen, Iglesias, Goldstein, & Bedore, 2014) was administered to each participant of the study. Children who scored below the 30th percentile were classified as having language difficulty. The reading dynamic assessment consisted of a pretest, teaching, and posttest phase in which the children were taught reading strategies and were asked to read nonsense words. In the teaching phase the students learned how to recode the same nonsense words tested in the pretest phase by using an onset-rime, analogous method. In the posttest phase the children read the same words used in the pretest and teaching phases. Each child was scored using a phoneme gain score and residuum gain score. The strategies each child used to read words was assessed and given a decoding strategy score. Lastly, each child's response to instruction was scored. The decoding strategy scores and response to instruction scores were combined to create a modifiability score. Interrater agreement on the scoring of response to instruction and number of correct sounds and words was 97%. Interrater agreement on the scoring reading strategy was 98%. In addition to the reading dynamic assessment, the children's reading abilities were also assessed using subtest from the DIBELS standardized assessment. *Results:* The dynamic assessment modifiability score had the highest correlation with first grade reading outcomes. The residuum gain score also correlated with first grade reading outcomes. However, the sounds gain scores shows no correlation with the reading outcome measures. Overall, the modifiability measures of the dynamic assessment had high evidence of validity in predicting which kindergarten students were susceptible to literacy challenges in first grade. The modifiability score had 100% sensitivity and 80% specificity for predicting oral reading fluency, 100% sensitivity and 88% specificity for predicting word identification, and 86% sensitivity and 85% specificity for predicting nonsense word fluency scores in first-grade. In comparison, the static measures used to assess the children's literacy resulted in a high over-classification of students as at risk for literacy difficulty. *Relevance to current work:* This study supports the assertion that dynamic assessment measures have a higher classification accuracy than static measures, even static measures designed for bilingual students. Static measures of reading result in the over-classification of bilingual students. In this study, modifiability scores had the highest sensitivity and specificity.

Restrepo, M. A., & Silverman, S. W. (2001). Validity of the Spanish Preschool Language Scale-3 for use with bilingual children. *American Journal of Speech-Language Pathology, 10,* 382-393.

Purpose: This purpose of this study was to investigate the validity of the Spanish Preschool Language Scale (SPLS-3; Zimmerman, Steiner, & Pond, 1993) In identifying language impairment in Spanish-speaking students. *Method:* The two researchers independently examined the SPLS-3 manual and assessed whether or not it met psychometric standards for validity. The manual was assessed using 15 criteria from McCauley and Swisher (1984) and Hutchinson (1996). In addition to these 15 criteria, the researchers also examined how representative the normed sample was of local bilingual children and the construct, content, concurrent, and criterion evidence of validity of the

SPLS-3. Thirty-seven bilingual, kindergarten children participated in the study. The children were balanced bilingual or Spanish-dominant speakers of varying language ability. The students were considered at risk for language impairment or typically developing based on teacher report. Each child was assessed using 5 measures: a Spanish language sample, a parent interview, the SPLS-3, a criterion-referenced receptive measure (CRR), a criterion-referenced expressive measure (CRE), and the Preschool Language Assessment Scale (PreLAS; Duncan & DeAvila, 1986). The parent interview was used to assess the language skills of teach child. The PreLAS was used to determine if each child's Spanish proficiency. The CRR used probed each child's understanding of grammatical morphemes in Spanish while the CRE measured the child's ability to use the same grammatical morphemes. All four measures (the PreLAs, CRR, CRE, and parent report) were used to compare with each child's performance on the SPLS-3. Reliability for the SPLS-3 was 94.7% and 98% for the parent interviews. *Results:* The results of this study showed that the information provided by the SPLS-3 manual provided insufficient evidence of concurrent validity, predictive validity, test-retest reliability, and construct validity. Results from the participants' scores on the SPLS-3 showed that 51% of the students scored more than one standard deviation below the mean when compared to the normative data of Spanish speakers provided by the SPLS-3. This indicates that the normative data provided by the SPLS-3 was likely not representative of the local bilingual children in the study and resulted in an over-classification of children. In the assessment of construct- and content-related evidence of validity, researchers found that several of the test items contained vocabulary and syntactical elements that were developmentally and culturally inappropriate for young Spanish-speaking children.

Additionally, when determining the criterion evidence of validity of the SPLS-3, researchers discovered that the receptive measures of the SPLS-3 correlated with the CRR, but the expressive measures did not correlate with the CRE. The total score of the SPLS-3 did not correlate with the total score of the participants on the PreLAS. Lastly, the results from the parent interviews showed that six children were considered at risk for language problems. However, of these 6 children, only one scored below one standard deviation from the mean indicating a lack of sensitivity. *Relevance to current work:* This study supports the assertion that static, norm-referenced tests, even those designed for Spanish-speakers can lack sensitivity and specificity and contain test items that are often culturally inappropriate. For these reasons, norm-referenced tests can be unreliable and poor measures of language impairment.

Spencer, T. D., Petersen, D. B., & Adams, J. L. (2015). Tier 2 language intervention for diverse preschoolers: An early-stage randomized control group study following an analysis of response to intervention. *American Journal of Speech-Language Pathology*, *24*, 619-636. doi: 10.1044/2015_AJSLP-14-0101

Objective: The purpose of this study was two-fold: First, to investigate how valid a dynamic assessment of narrative language was in identifying preschoolers in need of RTI Tier 2 intervention, and second, to determine the efficacy of using *Story Champs* language intervention in improving Tier 2 participants' narrative language when compared to a Tier 2 control group. *Method:* The study was divided into two phases. In the first phase a dynamic assessment of narrative language was used to identify participants in need of Tier 2 language support. The participants involved in the first phase were 41 children drawn from three different Head Start classes. The racial/ethnic profile of the 41 participants consisted of Latino, Native American, White, African American, and mixed ethnic backgrounds. The dynamic assessment followed a test-teach-retest format. The Narrative Language Measure (NLM; Peterson & Spencer, 2012) was used to assess the students during the pretest and posttest phases. In these two phases the participants told three story retells and were scored based on their inclusion of story grammar, episodic elements, and language complexity. Pre- and posttesting took approximately 3-5 minutes to complete, and scoring of the narratives elicited during these sessions took 4-5 minutes. The teaching phase of the dynamic assessment occurred in a large group setting over a period of three days. The child participants participated in three teaching sessions that each lasted 15-20 minutes. During the teaching sessions the examiner first modeled a narrative with pictures and then instructed the children to label each part of the narrative. The examiner then retold the story while the children produced gestures associated with the story. The examiner then had the children answer questions about the story. Lastly, the children were paired together and practiced retelling the story with each other. Students who scored below an 8 on the posttest (n=22) were selected to participant in the second research phase. The twenty-two participants were divided into a treatment (n=11, due to one drop out) and control group (n=10). The treatment group received Tier 2 language intervention following the *Story Champs* format across 18 sessions, two times a week over a period of 9 weeks. In addition to participating in the Story Champs intervention, participants were assessed using The Renfrew Bus Story and also produced personal stories in informal settings. Both the Renfrew Bus Story and personal stories were scored by the examiners. The participants' parents also helped their children complete take-home activities and completed a survey about the home activities. *Results:* Results of Phase 1, the dynamic assessment portion of the study, indicated that 17% of the children required no further intervention, 24% of students made gains from pre- to posttest, and 42% responded to the narrative teaching phase as evidenced by the small amount of gain from pre- to post. Following intervention in Phase 2, the treatment group had significantly higher scores on the NLM posttest, on follow-up measures, and the Renfrew Bus Story. However, the treatment and control group showed little difference in personal story scores. *Relevance to current work:* This study demonstrates that narrative based dynamic assessments can be used with young children and in a group setting. This study also evidences that dynamic assessment can be used to identify bilingual children who may benefit from Tier 2 support.

Squires, K. E., Lugo-Neris, M., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language & Communication Disorders, 49*, 60-74. doi: 10.1111/1460-6984.12044

Objective: The goal of this study was to perform a longitudinal study of bilingual Spanish-English speaking children in order to determine whether or not bilingual children with LI differ from their typically developing peers in macro and micro narrative structures from kindergarten to first grade. *Methods:* Twenty-one bilingual children with LI participated in the study. In order to initially identify the children with LI, three bilingual SLPs used a 6-

point scale to rate the children’s performance in Spanish and English vocabulary, morphosyntax, and narration. Students who scored at or below a 2 in both languages were considered to have LI. Each of the 21 children with LI were matched to a bilingual peer based on sex, age, month of birth, IQ score, and English and Spanish language exposure. The total number of participants was 42. During narrative assessment, each child was modeled a narrative retell using a wordless picture book and was then asked to retell that same wordless picture book. Each child provided both an English and Spanish retelling once during their kindergarten year and then again in their first-grade year. During both the kindergarten assessment and first-grade assessment, each participant retold the same story in the same language. Participants’ retells were scored in terms of inclusion of macrostructure and microstructure elements using a 0-3 point scale with 3 being the highest score. Interrater reliability measures for scoring were 96%. *Results:* In terms of macrostructure, the TD group had overall higher scores than the LI group. Both groups improved in macrostructure from kindergarten to first grade. However, the TD group made much larger gains in macrostructure scores from kindergarten to first grade than the LI group. In addition, language exposure to English influenced macrostructure scores. Participants with less English experience scored better in Spanish and lower in their English narrative retells in terms of macrostructure. Students with more English exposure performed better in their English retells and performed more poorly in their Spanish retells. In terms of microstructure, the TD group had higher scores than the LI group. Similar to the macrostructure findings, students with more English exposure had higher English microstructure scores but lower Spanish microstructure scores. Interestingly, the TD group showed significant improvement in microstructure scores in Spanish from kindergarten to first-grade but their English microstructure scores decreased from kindergarten to first grade. The LI group showed no improvement in microstructure scores between kindergarten and first grade in either Spanish or English. These results indicate that bilinguals can learn to easily transfer content-based information into their L2 languages over time, but that they must learn to independently transfer microstructure linguistic elements into their L2. For bilinguals with LI, this transfer is even more difficult than for their typical peers as they require more exposure to macro and microstructure linguistic concepts before they can successfully transfer them into their L2. *Relevance to current work:* This study supports the assertion that low narrative language skills correlate with language impairment across languages. This study also shows that children who are Spanish dominant speakers will have stronger narratives in Spanish than in English. Additionally, the results of this study indicated that some aspects of narrative language are difficult to learn in a child's L2 even if they have typically developing language.

Ukrainetz, T. A., Stacey, H., Walsh, C., & Coyle, C. (2000). A preliminary investigation of dynamic assessment with Native American kindergarteners. *Language, Speech, and Hearing Services in Schools, 31*, 142-154.

Purpose: Ukrainetz, Walsh, & Coyle (2000) used a test-teach-retest dynamic assessment format to determine if a dynamic assessment was a more culturally appropriate measure of language ability for 23 Arapahoe and Shoshone kindergarten children than standardized, norm-referenced testing. *Methods:* Fifteen of the 23 students were considered "normal language learners" while eight were considered "weak language learners" based on teacher report and researcher observation. All 23 participants were primarily English-only language speakers. Participants were assessed using a test-teach-retest format over a period of three weeks. Pre- and posttests each took 20 minutes to administer. The teaching phase consisted of two, 30-minute sessions, with two participants present during each teaching phase. During the two teaching phases, students were taught categorization skills by learning to group similar words under a unifying category. Following the teaching phases, the examiner scored each student's learning strategies and responsiveness to intervention using two 5-point Likert scales. Expressive and receptive subtests taken from the Assessing Semantic Skills through Everyday Themes (ASSETS; Barrett, 1988), a standardized, norm-referenced test, were used as pre- and posttest measures to investigate the effect of the mediation sessions. The ASSETS was administered 1-5 days prior to administration of the dynamic assessment and again 1-5 days after the administration of the dynamic assessment. Interrater reliability in scoring modifiability (learning strategies and responsiveness to intervention) was 94%. Interrater reliability on the scoring of the ASSETS was 96%. *Results:* Results of the study showed that modifiability scores and scores from the second administration of the ASSETS (posttest) were higher for the "normal language learners" than the scores of the "weaker language learners." The study also found that the responsiveness of the child during the teaching phase was a better predictor of language learning difficulty than the students' learning strategies used in the teaching phase. The sensitivity and specificity of the dynamic assessment was not reported. *Relevance to current work:* This study demonstrates that a categorization dynamic assessment may be an effective tool to differentiate language difference from disorder in culturally diverse, monolingual English speakers, and supports the assertion that dynamic assessments may be more valid assessment measures than static measures. However, because sensitivity and specificity were not reported, the classification accuracy of this assessment is unknown.