



Theses and Dissertations

2020-07-13

The Classification Accuracy of a Dynamic Assessment of Language in Culturally and Linguistically Diverse Children When Using Response to Intervention as a Measure of Language Ability

Yuberkys Fryer
Brigham Young University

Follow this and additional works at: <https://scholarsarchive.byu.edu/etd>



Part of the [Education Commons](#)

BYU ScholarsArchive Citation

Fryer, Yuberkys, "The Classification Accuracy of a Dynamic Assessment of Language in Culturally and Linguistically Diverse Children When Using Response to Intervention as a Measure of Language Ability" (2020). *Theses and Dissertations*. 9146.

<https://scholarsarchive.byu.edu/etd/9146>

This Thesis is brought to you for free and open access by BYU ScholarsArchive. It has been accepted for inclusion in Theses and Dissertations by an authorized administrator of BYU ScholarsArchive. For more information, please contact ellen_amatangelo@byu.edu.

The Classification Accuracy of a Dynamic Assessment of Language in Culturally and
Linguistically Diverse Children When Using Response to Intervention
as a Measure of Language Ability

Yuberkys Fryer

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
Shawn L. Nissen
Christopher Dromey

Department of Communication Disorders
Brigham Young University

Copyright © 2020 Yuberkys Fryer

All Rights Reserved

ABSTRACT

The Classification Accuracy of a Dynamic Assessment of Language in Culturally and Linguistically Diverse Children When Using Response to Intervention as a Measure of Language Ability

Yuberkys Fryer

Department of Communication Disorders, BYU
Master of Science

The purpose of this study was to examine the extent to which modifiability ratings and gains in narrative language, made through intervention over time with culturally and linguistically diverse children, aligned with the results of a diagnostic dynamic assessment of language. This study also examined the sensitivity and specificity of the dynamic assessment when response to language intervention was used as the primary indicator of language disorder (LD). A total of 32 culturally and linguistically diverse students from an elementary school in Utah participated in this study, with 17 students with LD and 15 students without LD. Students were administered a dynamic assessment of language and were then provided small group narrative-based language intervention for several weeks. Student progress was monitored each week by collecting narrative language samples. Modifiability ratings were also collected, which provided information on student learning potential. Progress monitoring gain scores from the first intervention session to the last intervention session and mean modifiability ratings were compared between children with and without language disorder. Logistic regression and receiver operator characteristic analyses were conducted to obtain classification accuracy information. The results of this study indicated that growth in narrative language due to intervention did not reflect the results of the dynamic assessment; however, modifiability scores, which measure a student's difficulty in learning language, aligned with the dynamic assessment results. Sensitivity was 94% and specificity was 71%. It is possible that a dynamic assessment of language may be a less biased approach to diagnose LD in culturally and linguistically diverse students.

Keywords: dynamic assessment, language disorder, response to intervention, narrative

ACKNOWLEDGMENTS

The author would like to thank the following committee members for their meaningful contribution: Dr. Douglas B. Petersen, Dr. Shawn L. Nissen, and Dr. Christopher Dromey. Their expertise, guidance, and experience were vital to the success of this project. The author also thanks all members of the DYMOND team for their time and effort collecting and analyzing the data used in this project.

TABLE OF CONTENTS

TITLE PAGE.....	i
ABSTRACT	ii
ACKNOWLEDGMENTS.....	iii
TABLE OF CONTENTS.....	iv
LIST OF TABLES.....	vi
LIST OF FIGURES	vii
DESCRIPTION OF THESIS STRUCTURE AND CONTENT.....	viii
Introduction	1
Norm-Referenced Tests and Poor Classification Accuracy	3
Dynamic Assessment	4
Method	9
Participants	9
Procedures	10
Measures.....	12
DYMOND Dynamic Assessment of Language	12
Dynamic Assessment Pretest	12
Dynamic Assessment Teaching Phase.....	13
Dynamic Assessment Modifiability	14
Dynamic Assessment Posttest.....	14
CUBED: Narrative Language Measures: Listening (NLM).....	14
Intervention Procedures.....	16

Small Group Story Champs	16
Interventionists and Fidelity of Intervention.....	16
Test Administration Fidelity and Scoring Reliability	16
Results	17
Research Question 1	17
NLM Gain Scores from Intervention Session 1 to the Last Intervention Session	18
Modifiability Scores from Each Intervention Session.....	18
Mean Modifiability Total Scores	19
Research Question 2.....	19
Discussion	20
Narrative Language Measure Analysis	21
Modifiability Total Score	22
Sensitivity and Specificity	22
Limitations and Future Research	23
Conclusions.....	23
References	25
APPENDIX A: Annotated Bibliography	30
APPENDIX B: CUBED Narrative Language Measures: Listening (NLM: Listening)	45
APPENDIX C: Small Group Narrative Intervention Fidelity Checklist	46
APPENDIX D: IRB Approval Form	47

LIST OF TABLES

Table 1	<i>Demographic Information for all Participants</i>	10
Table 2	<i>Number of Intervention Sessions Received by Each Group</i>	11
Table 3	<i>Means and Standard Deviations of Predictor Variables</i>	17

LIST OF FIGURES

<i>Figure 1.</i>	Level and Slope From the First NLM Administration to the Last NLM Administration.	18
<i>Figure 2.</i>	Area Under the Curve (AUC) Output Indicating Optimal Balance of Sensitivity and Specificity.	20

DESCRIPTION OF THESIS STRUCTURE AND CONTENT

To adhere to traditional thesis requirements and journal publication formats, this thesis, *An Examination of the Classification Accuracy of a Dynamic Assessment of Language when using Response to Intervention as a Measure of Language Ability*, is written in a hybrid format. This thesis is part of a larger study on Dynamic Assessment. The initial pages of the thesis adhere to university requirements while the thesis report is presented in journal article format. The annotated bibliography is included in Appendix A. Appendix B is the *CUBED* Narrative Language Measures (*NLM*), followed by Appendix C, which includes the *Story Champs* Small Group Narrative Intervention Fidelity Checklist. Appendix D contains the IRB approval form.

Introduction

The Hispanic and Latino population is the largest ethnic minority in the United States. According to the U.S. Census in 2016, there were 58.9 million Hispanics in the United States (U.S. Census Bureau, 2018), that is, 18.1% of the total population, with 40 million people speaking Spanish at home in the United States (U.S. Census Bureau, 2018). Moreover, it is projected that the Hispanic population will continue to grow in the United States. It is estimated that in 2060, the Spanish-speaking Hispanic population will account for 28.6% of the total population, that will result in 119 million Hispanic individuals living in the United States (Colby & Ortman, 2015). In addition, the student Hispanic population will continue to expand and grow at a rapid rate. According to Bauman (2017), from 1996 to 2016, the number of Hispanic students enrolled in school, colleges and universities doubled from 8.8 million to 17.9 million, representing 22.7% of all individuals enrolled in school. According to the latest report for the National Center for Education Statistics (NCES), Spanish was the home language of 3.79 million English Language Learners in public schools in the United States (2019).

Furthermore, the number of Spanish speakers is not only growing across the United States, but also across the world. In fact, the latest annual report from the Instituto Cervantes (2018), Spanish is the second-most spoken language with more than 577 million people speaking Spanish around the world. By 2050, the Spanish speaking population is predicted to increase to 756 million, making it the most widely spoken language across the world.

The Spanish-speaking school-age children within the United States and across the world have different language and life experiences. Many Spanish-speaking children across the world are multilingual, learning Spanish as either their first or second (or third) language. For instance, in Guatemala, Mexico, Belize and Honduras, over 30 Mayan languages are commonly

spoken at home (Kaufman, 1974). Many of these Mayan-speaking children learn one of the Mayan languages from their parents before they learn Spanish in school. These children often learn Spanish sequentially, and this second language may have weaker vocabulary and syntax than their first language. Also, these children are often living in underdeveloped areas and do not regularly attend elementary or secondary schools. Furthermore, in 2008, 88% of elementary schools and 93% of secondary schools in the United States with foreign language programs offered Spanish (Wiley et al., 2014). These Spanish-speaking children have different language experiences when learning Spanish as a second/third language. In addition, most of these Spanish-speaking, English language learners in the U.S. live in poverty and have a lack of healthcare (National Council on Disability, 2018). These Hispanic families from lower Social Economic Status (SES) backgrounds tend to talk less to their children, use limited vocabulary, and are less likely to read to their children compared to monolingual, English-speaking families (Sonnenschein et al., 2017). Therefore, the proficiency and dominance in Spanish varies among these multilingual children.

Because these multilingual students are at varying stages of language learning and attrition (Restrepo & Kruth, 2000), norm-referenced tests (NRTs) have poor evidence of being able to differentiate between multi-lingual children who have a language disorder (LD) and multi-lingual children who are in varying stages of language learning and who do not have a disorder (Williams & McLeod, 2012). Today, NRTs are the most common tool used in the United States to identify a LD. Yet, NRTs have generally yielded poor specificity and sensitivity in identifying LD in school-aged children (Spaulding et al., 2006). Sensitivity is the capability of a test to accurately identify children with a LD Specificity, on the other hand, is the capability of

a test to accurately identify children without a LD. To have enough evidence of classification validity, diagnostic tests should yield sensitivity and specificity levels that are at or above 80% (Spaulding et al., 2006).

Norm-Referenced Tests and Poor Classification Accuracy

The classification accuracy of NRTs when used with culturally and linguistically diverse (CLD) children is often weaker. Laing and Kamhi (2003) explained that NRTs in the English language are biased and culturally inappropriate when used with CLD children due to three main issues: content bias, linguistic bias, and disproportionate representation in normative samples. Content bias occurs when it is assumed that all children have been exposed to identical concepts and vocabulary or similar life experiences as the mainstream culture. Linguistic bias occurs when inconsistencies exist between a) the examiner's language or dialect, b) the child's language or dialect, and c) the child's language or dialect expectations in his/her response. Linguistic bias can be problematic in the context of NRTs because it can classify a child as atypical when the child might have typical language development in their dominant language or dialect. NRTs can also be biased when disproportionate representation in normative samples occurs. Culturally and Linguistically Diverse (CLD) students have often been excluded from the normative samples used to create norms. Despite the fact that most tests now include different ethnicities in their sample data to better represent the diversity of students in the U.S. schools, CLD students with and without LD are often underrepresented by NRTs (Laing & Kamhi, 2003).

Although English NRTs of language are often biased and unsuited to assess culturally and linguistically diverse students, using NRTs that are suited to the student's native language has also failed to yield appropriate levels of sensitivity and specificity. For example, two of the NRTs commonly used are the Spanish Preschool Language Scale (SPLS-3; Zimmerman et al.,

1993) and the Clinical Evaluation of Language Fundamentals Fourth Edition-Spanish Version (CELF-4S; Semel, Wiig, & Secord, 2006). Restrepo and Silverman (2001) evaluated the validity of the SPLS-3 in identifying Spanish-speaking children with LD. They found that 51% of typically developing (TD) children from a local sample were more than one standard deviation below the mean of the normative data of Spanish children in the SPLS-3. They also found that most of the test items were culturally inappropriate. Another study by Barragan et al. (2018) examined the performance of 656 Spanish-speaking dual-language learners, ages 5;0-7;11, on the Clinical Evaluation of Language Fundamentals Fourth Edition, Spanish (CELF-4S; Semel et al., 2006). These children were tested for LD using the CELF-4S and the English Structured Photographic Expressive Language Test (Dawson et al., 2003). The study found that the CELF4S overidentified low-income Spanish-English dual-language learners attending English only schools. The CELF-4S manual suggests that the cut off standard score for core language should be 85, which is one standard deviation below the mean. When using this cutoff score, the test yielded a sensitivity score of 94% and a specificity score of 65%. The last two studies (Restrepo & Silverman, 2001) and (Barragan et al., 2018) showed that NRTs designed to evaluate Spanish speakers often lack adequate sensitivity and specificity and may result in over or under classification of LD in Spanish-speaking children.

Dynamic Assessment

Dynamic assessment is an alternative to NRTs that could reduce the bias that is often found in traditional, static language assessments when assessing bilingual children. Dynamic Assessment and NRTs differ in that dynamic assessments are measures of students' learning abilities rather than single static measures of a child's knowledge at a given point in time. Intervention is part of the assessment as the clinician teaches a concept and provides support

during the learning process. Depending on the child's performance during the intervention, the examiner draws conclusions about the child's gains during the intervention, the amount of prompts the child needed to be successful, and the strategies the child used (Kapantzoglou et al., 2012). Dynamic assessment draws from the principles that measure student learning, often using a pretest-teach-retest model. Vygotsky's Zone of Proximal Development (ZPD; 1978) suggests that a child's zone of proximal development focuses on what tasks a student can successfully perform independently and what tasks are outside of a student's ability to accomplish at their current level of functioning. The pretest-teach-test model allows the examiner to determine how well a child can learn with direct instruction. Having this information will help identify the student's ZPD during the teaching phases, where Mediated Learning Experiences are provided (MLE; Feuerstein, 1979). During the teaching phase, individual instruction is provided to determine the student's learning potential, or modifiability, which is a measure of how much effort is required by the examiner to help the child learn and make progress during the MLE sessions.

Several studies have investigated dynamic assessment of language. For example, Peña and Iglesias (1992), compared the accuracy of a dynamic assessment of vocabulary against a standardized vocabulary assessment in identifying culturally diverse children with LD. A total of 50 African American and Puerto Rican students from three Head Start classes in Northern Philadelphia participated in this study. All of the students were exposed to English and Spanish in the classroom. Two standardized test instruments were used: The Expressive One-Word Picture Vocabulary (EOWPVT; Gardner, 1979) and the Comprehension subtest of the Stanford-Binet Intelligence Scale (CSSB; Thorndike et al., 1986).

Those students who scored low on EOWPVT received mediation training (a dynamic assessment teaching phase). The mediation consisted of two 20-minute sessions that focused on improving vocabulary labeling abilities of the students. After each mediation session, the clinician scored each student based on their responsiveness, examiner effort, and transfer of skills to obtain an overall modifiability rating. After the two mediation sessions, the students were assessed using the EOWPVT. The results of this study indicated that both TD students and students with language disabilities scored equally low on the EOWPVT during the pretest and that classification accuracy of the dynamic assessment was 92% of the LD cases. Finally, the TD children had higher modifiability scores as well as higher gains than the students with LD.

Peña et al. (2006) used dynamic assessment to identify which variables in a dynamic assessment of narrative language were most predictive of LD. They administered the dynamic assessment to 71 first and second grade diverse students from central Texas. In this study, children were from different backgrounds including African American, European American and Latino American. The dynamic assessment consisted of a pretest-teach-test model. Participants were divided into three different groups: a control group that consisted of 30 children, a typical developing group that consisted of 27 children, and a language impaired group that consisted of 14 children. During pretest and posttest, all children told a story based on two different wordless picture books. Children in the TD group and language impaired group received two individualized 30-minute sessions focusing on narrative skills and strategies. At the end of the second intervention session, examiners evaluated how much support was required based on 5-point Likert scale. Each child's responsivity was also evaluated on 5-point Likert scale. A score of 5 meant high child responsivity and a score of 1 meant low child responsivity. The entire dynamic assessment took several hours to complete across multiple days. The results of this

research indicated that modifiability and posttests scores provided 100% sensitivity and 100% specificity.

Kramer et al. (2009) conducted a study to probe the accuracy of the dynamic assessment of narrative language that Peña et al. (2006) studied. In this study, the dynamic assessment was administered to a group of third-grade children from Samson Cree Nation Reserve in Alberta, Canada. The dynamic assessment was administered to 17 children; 5 of them were labeled as having LD and 12 of them were classified as having typical language development (TD). The administration of the entire dynamic assessment was finished in a period of 4 days. In this study they used the same wordless picture books as well as the Likert scoring scale to measure each student's modifiability, responsiveness, and narrative production as the Dynamic Assessment and Intervention (DAI). The narrative transcripts of the Dynamic Assessment were scored by two examiners. The final scoring decisions of pretest, posttest, and modifiability were reached through consensus between the two examiners. However, interrater reliability on modifiability scoring and pre and posttest scores were not reported. The results of this study indicated that the dynamic assessment was accurate for children in third-grade because it demonstrated classification accuracy in identifying children with LD. Although both groups had similar scores at the pretest phase, typical language students made greater improvements in targeted and nontargeted narrative elements. Modifiability ratings and posttest performance most accurately classified students, yielding 100% sensitivity and 92% specificity. This study also indicated that modifiability alone yielded 100% sensitivity, yet with only 75% specificity.

Recently, researchers investigated the classification accuracy of an English narrative dynamic assessment for identifying LD in Spanish-English bilingual kindergarten to third-grade students (Petersen et al., 2017). The study used a more concise dynamic assessment with a

realtime scoring procedure to determine whether LD could be identified in less time than conventional dynamic assessment measures as long as appropriate classification accuracy was maintained. The study included 42 Hispanic children who were bilingual in both English and Spanish (10 with LD and 32 without LD) from a large urban school district in the mountain west. The students were classified as balanced bilingual, Spanish dominant or English dominant. The dynamic assessment consisted of two 25-minute test-teach-retest sessions. Each session consisted of a pretest narrative retell, a narrative retell teaching phase, and posttest narrative retell. Both the pretest and posttest narrative retell and modifiability ratings were scored during the session. During the teaching phase, clinicians individually targeted story grammar and adverbial subordinate clauses. The pre and posttests of the dynamic assessment were scored based on (a) the nine story grammar elements (i.e., character, setting, problem, emotion, plan, attempt, consequence, ending and ending emotion), (b) occurrence of conjunctions (i.e., then, when, because, and after), and (c) complexity of episodic structure. The teaching phase targeted each of the elements used in the pre and posttest. After each teaching phase, the examiner scored the children using a modifiability rating scale used in previous dynamic assessment research. The results of this study yielded high classification accuracy. The overall modifiability from both dynamic assessments sessions yielded 100% specificity and 100% sensitivity. In addition, the modifiability score for one of the 25-minute sessions yielded to 100% sensitivity and 91% specificity.

Although the evidence to support dynamic assessment of English is promising, more research is needed to investigate whether a dynamic assessment will accurately identify culturally and linguistically diverse students who have a language learning disorder. Therefore, the purpose of this study was to examine if the results of a dynamic assessment align with

modifiability ratings and gains in narrative language made through intervention over time and whether sensitivity and specificity are adequate when this response to intervention is used as the primary indicator of LD. The research questions were as follows:

1. Do students identified as having typical language using a dynamic assessment make stronger gains over time on the *NLM* and have higher modifiability scores when small group language intervention is provided than children identified as having a LD using a dynamic assessment?
2. To what extent does dynamic assessment of language in English accurately identify school-age diverse children with LD (sensitivity) and without LD (specificity) when response to evidence-based English language intervention over time is used to diagnose LD?

Method

Participants

The BYU Institutional Review Board approved this study. Participants for this study were recruited from an elementary school in Utah. All first through sixth-grade students in the school were invited to participate. Two-hundred and nine children had parent/guardian consent to participate and were included in this study. Of those 209 children, 27 were identified as having a LD in the spring using the English Dynamic Assessment of Oral Narrative Discourse (DYMOND; Petersen et al., 2017). Of those 27 identified as having a LD in the spring, 17 were still in the school in the fall. Those 17 children identified by the DYMOND as having a LD were matched to 17 children with typical language development; Yet, two of those matching TD students moved early in the fall and were not able to participate in the study. Thus, there were a total of 32 students who participated in this study, with 17 having a LD and 15 with typical

language development. All students were matched by grade, gender, and whether the student was bilingual or monolingual to the fullest extent possible. Table 1 provides descriptive information about each participant, including information on bilingual English/Spanish status, ethnicity/race, gender, grade, and whether the child has an Individualized Education Program (IEP) for language services.

Table 1

Demographic Information for all Participants

	<u>Language Disorder</u>	<u>Typically Developing</u>
Gender		
Female	<i>n</i> = 5 (29%)	<i>n</i> = 7 (47%)
Male	<i>n</i> = 12 (58%)	<i>n</i> = 8 (42%)
Ethnicity		
Caucasian	<i>n</i> = 9 (44%)	<i>n</i> = 8 (56%)
Hispanic	<i>n</i> = 8 (62%)	<i>n</i> = 7 (38%)
Bilingual English/Spanish	<i>n</i> = 5 (29%)	<i>n</i> = 6 (40%)
Grade		
Second	<i>n</i> = 8	<i>n</i> = 8
Third	<i>n</i> = 5	<i>n</i> = 5
Fourth	<i>n</i> = 1	<i>n</i> = 0
Fifth	<i>n</i> = 1	<i>n</i> = 0
Sixth	<i>n</i> = 2	<i>n</i> = 2

Procedures

After administering the DYMOND in the spring, the children with LD and the children with typical language development received *Story Champs* intervention in small groups in English in the fall of the following school year. Children with a LD as identified by the

DYMOND were placed into groups of two or three students. Matching students were placed in equivalent sized groups and received the same dosage of intervention. There were 12 small groups, and each group dyad received anywhere from one intervention session to six intervention sessions across 4 weeks. A detailed intervention schedule in Table 2. Each week on different day from when intervention was delivered, students' progress in narrative language was assessed by blinded research assistants using one *NLM*: Listening Parallel form. Immediately at the end of each intervention session, research assistants completed a modifiability form that reflected the examiner's effort to conduct the intervention and the children's responsiveness to the intervention.

Table 2

Number of Intervention Sessions Received by Each Group

	Number of Sessions					
	1	2	3	4	5	6
LD Group 1 (Second Grade)	Green	Green	Yellow	Yellow		
LD Group 2 (Second Grade) LD	Green	Green	Green	Green	Green	
Group 3 (Second Grade) TD	Green	Green				
Group 4 (Second Grade) TD	Green	Green	Green			
Group 5 (Second Grade)	Green	Green	Green	Green		
TD Group 6 (Second Grade)	Green	Green	Green	Green		
LD Group 7 (Third Grade)	Green	Yellow	Yellow			
TD Group 8 (Third Grade) TD	Green	Green	Green	Green	Green	Yellow
Group 9 (Third Grade)	Green	Green	Green	Yellow	Yellow	
LD Group 10 (Third/Fourth Grade)	Green	Green	Green	Green		
LD Group 11 (Fifth/Sixth Grade) TD	Green	Green	Green	Yellow	Yellow	Yellow
Group 12 (Fifth/Sixth Grade)	Green	Green	Green	Green	Yellow	Yellow
	Green		Yellow			

Note. *LD* = Language Disorder; *TD* = Typical Language Development; Green blocks = Sessions where all students in the group received intervention. Yellow blocks = Sessions where one or more students in the group were absent.

Three primary sources of evidence were used to diagnose LD. A student had to meet all three criteria to be correctly classified. First, progress over time in response to evidence-based language intervention was examined. It was hypothesized that the students who had a LD would make slower progress than matched TD students. Second, the mean of the modifiability ratings collected at the end of each intervention session was analyzed. It was hypothesized that students with a LD would have lower mean modifiability ratings than TD students. Third, students with a LD either had an active IEP for language or there were educator concerns about the student's language. Conversely, students without a LD did not have an IEP, and educators were not concerned about the student's academic skills. In order to determine whether the DYMOND accurately identified TD, the results of the DYMOND were compared to the students' definitive language ability classification per the criteria outlined.

Measures

DYMOND Dynamic Assessment of Language

All the children who participated in this study were given the dynamic assessment of language. The dynamic assessment of language includes four steps: a pretest, a teaching phase, a modifiability rating scale, and a posttest. The dynamic assessment of language took about ten minutes, depending on the child's responsiveness.

Dynamic Assessment Pretest

The pretest involved the examiner reading a brief narrative (story) and having the student retell that narrative. The students were assessed on their inclusion of story grammar elements and elements of language complexity (e.g., *because*, *when*, *after*). The stories were scored in realtime using a point system. Each retell had a maximum score of 35 points. This maximum score was comprised of the story grammar subtotal and the language complexity scores. Two points were

awarded for the inclusion of each story grammar element, which produced a maximum total of 26 points. One point (up to 9 points) was given each time the student used the subordinating conjunctions *because*, *when*, or *after*.

Dynamic Assessment Teaching Phase

The teaching phase consisted of two steps which were designed to help the children learn to independently produce complete narrative episodes (i.e., including at least the problem, attempt, consequence, and ending) and improve their language complexity. In the first step, a set of pictures with corresponding story grammar icons were placed in front of the child. The examiner retold the pretest story while simultaneously pointing to the corresponding pictures and explicitly teaching icons which represented important story grammar elements (e.g., “This is how Sam felt. He was sad.”). Following this part of the instruction, the child used the pictures and icons to retell the story, and the examiner helped the child include all story grammar elements and/or include language complexity targets. Once the child completed the retell with the pictures and icons, they moved on to the next step of the teaching phase. In the second step, the pictures were removed and the icons were left for the student to see. The student was then asked to retell the story again, using only the icons. The examiner again provided support and helped the child retell the story while including all appropriate story grammar elements and any language complexity targets.

An over-correction procedure was employed during both steps of teaching phase. If a student omitted or skipped a story grammar element, the examiner immediately stopped the student and provided a Level 1 prompt, which was an open-ended question. If the child did not respond to the open-ended prompt, the examiner provided a Level 2 prompt, which entailed modeling an appropriate response and having the student repeat it. Following either prompt, the

examiner instructed the child go back one step (story grammar element) and start telling the story from that point, including the missing story grammar element that time. In addition to focusing on teaching story grammar elements, the examiner was permitted to focus on increasing language complexity by prompting the use of the subordinating conjunctions such as *because*, *when*, or *after*. This focus on subordinating conjunctions typically only occurred if a student readily produced all of the story grammar elements.

Dynamic Assessment Modifiability

Immediately following the teaching phase, the examiner rated the student's modifiability (ability to learn) using a set of detailed modifiability rating scales. Using a 5-point scale, the examiner rated the student on the following criteria: response to prompts, degree of transfer, attention to teaching, ease of teaching, frustration, and disruptions. The examiner then totaled each score, with the potential to have a maximum score of 24. This score was defined as the total modifiability score. Then, the examiner rated the student on a scale of 0-4 on an overall scale, which reflected the final judgement score. A score of 4 represented relative ease in learning while a 0 represented difficulty learning.

Dynamic Assessment Posttest

The posttest followed the same procedure as the pretest, except with a different story of similar structure and complexity. The pretest and posttest stories were matched in language complexity (e.g., story length, use of tier-two words, dual-episode story structure, inclusion of subordinate clauses).

CUBED: Narrative Language Measures: Listening (NLM)

The *Narrative Language Measures: Listening (NLM)* subtest of the *CUBED* (Petersen & Spencer, 2012) that requires the retelling of a brief story was used as a progress monitoring tool.

Research assistants administered one grade appropriate, parallel form of the *NLM Listening* in English each week for six weeks to every student. The *NLM* is comprised of four sections that provide information on personal-themed narrative retells, personal story generations, story grammar comprehension, and inferential vocabulary comprehension. Only the narrative retell subtest from the *NLM* was administered and analyzed for this study. The *NLM* is a standardized, criterion-referenced general outcome measure with 25 parallel forms for each grade (pre-k to 3rd grade). The *NLM* is used to assess children's narrative language growth. It involves standardized administration and scoring procedures. The narrative retell subtest measures the comprehension and production of story grammar and limited aspects of complex language within personal themed narratives. Psychometric analyses indicate that the *NLM* has good to excellent reliability and validity (Petersen & Spencer, 2012).

To administer the *NLM*, research assistants read a model story, asked the child to retell it, and then listened to the child's story while providing only neutral prompts. Pictures were not used in the elicitation of the narrative retells. The *NLM* includes a scoring rubric designed to score student retells from each parallel story in real time. Stories were scored for the clarity and completeness of story grammar elements (character, setting, problem, feeling, action, consequence, and ending) on a 0-2 scale with weighted points for episodic elements (e.g., problem, action, consequence). Language complexity features such as the use of causal subordinating conjunctions (*because*) and temporal subordinating conjunctions (*after*, *when*) were scored for their frequency. Total *NLM* retell scores were calculated by summing the story grammar, language complexity, and episodic points. The time required for individual administration of each story was approximately 1-2 minutes.

Intervention Procedures

Small Group Story Champs

Research assistants used the *Story Champs* small group procedures with each small group of students one time per week for approximately 10-15 minutes per session. The small group intervention adhered to the small group procedures of *Story Champs* (Spencer & Petersen, 2012). The program includes multiple personal themed stories with accompanying pictures. Pictures were large enough to spread across a small table and allow for all children in the small group to see them. Additional visual materials included brightly colored story grammar icons representing the major parts of the story. Story games were used to increase children's active engagement while they listened to their peer tell a story individually. Materials for story games included small wooden sticks with the icons on them, small cubes with the icons on them, and bingo cards with the icons on them. Story gestures were also used in a game format, but materials were not required to play.

Interventionists and Fidelity of Intervention

Before serving as interventionists, the nine research assistants participated in a 4-hour training on the implementation of multi-tiered systems of language support using the *Story Champs* procedures. Research assistants practiced with each other, and received coaching and feedback from the lead researchers. Throughout the intervention phase, the researchers observed the research assistants conducting the intervention at least five times.

Test Administration Fidelity and Scoring Reliability

Prior to the study, the research assistants were trained in the administration and scoring of the *NLM Listening* for a minimum of 30 minutes. These research assistants administered and scored the narrative retells in real time and also audio recorded each assessment. Twenty percent

of the *NLM* retells from all assessment times were randomly selected to be scored by independent scorers. A large team of student research assistants independently listened to and scored the retells in real time. The following formula was used to calculate percent agreement: Number of agreements divided by agreements plus disagreements, multiplied by 100. For treatment fidelity, all intervention sessions were audio recorded and 20% of those sessions were randomly selected for a fidelity of test administration examination. An independent research assistant listened to each of the audio recordings and completed a multi-step fidelity checklist. For each one, the percent of steps completed correctly was calculated.

Results

Results are organized by research question. Means and standard deviations for the predictor variables are presented in Table 3.

Table 3

Means and Standard Deviations of Predictor Variables

	First Intervention NLM	Last Intervention NLM	NLM Intervention Gain	Mean Final Judgment Modifiability Score	Mean Total Modifiability Score
Language Disorder	11.18 (7.38)	13.82 (6.87)	2.65 (6.36)	3.28 (0.62)	19.31 (3.87)
Typical Language	16.38 (6.48)	17.33 (6.21)	1.08 (9.13)	3.79 (0.34)	21.78 (2.32) *

Note. *Statistically significant difference <.05.

Research Question 1

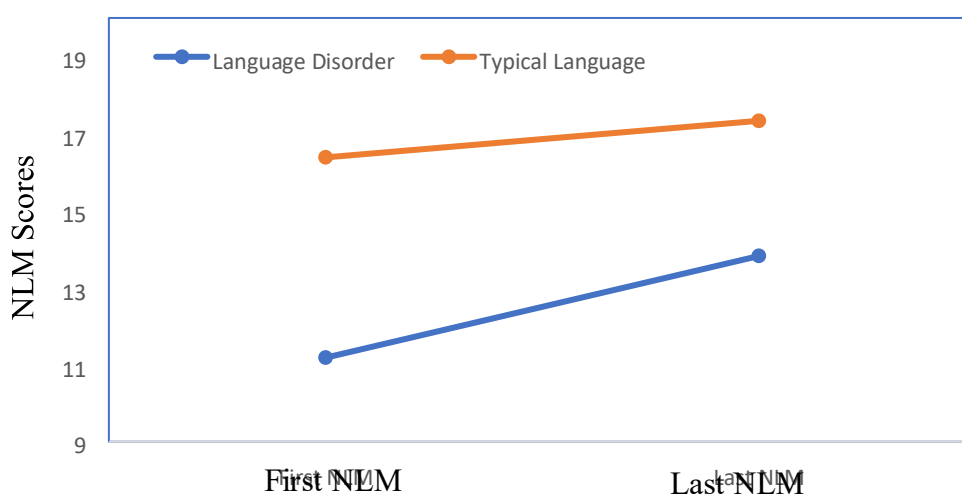
The first research question examined whether students with typical language would make stronger gains over time on the *NLM* and have higher modifiability scores than children with LD when small group language intervention is provided.

NLM Gain Scores from Intervention Session 1 to the Last Intervention Session

There was no significant difference between typical and LD groups for the gains from the first *NLM* administered during intervention to the last *NLM* administered during intervention (with TD mean = 2.65, *SD* = 6.36, with typical language development = 1.08, *SD* 9.13; $t = .67$ $p = .51$), see Figure 1.

Figure 1

Level and Slope From the First NLM Administration to the Last NLM Administration



Modifiability Scores from Each Intervention Session

Independent samples *t*-test were conducted to determine whether modifiability scores were significantly different between the students with LD and the students with typical language development according to the DYMond. For the total modifiability 1, there was no significant difference, (with LD mean = 17.76, *SD* = 5.39, typical language development mean = 20.64, *SD* 4.24; $t = 1.63$ $p = .11$). For the total modifiability 2, there was a significant difference, (with LD mean = 18.76, *SD* = 3.88, typical language development mean = 21.80, *SD* 3.17; $t = 2.43$, $p = .02$). For the total Modifiability 3, there was no significant difference, (with LD mean = 19.73,

SD = 5.06, typical language development mean = 21.33, SD 3.99; $t = .92$ $p = .37$). For the total Modifiability 4, there was no significant difference, (with LD mean=20.40, SD = 3.68, typical language development mean = 22.50, SD 1.20; $t = 1.56$ $p = .06$). For the total Modifiability 5, there were only two students with typical language development scored. For the total Modifiability 6, there were 0 students with typical language development scored.

Mean Modifiability Total Scores

Independent samples t -test were conducted to determine whether mean modifiability total scores were significantly different between the students with LD and the students with typically language development for LD according to the DYMOND. For the average total modifiability score, there was a significant difference, (with LD mean = 19.31, $SD = 3.87$, with typical language development mean = 21.78, $SD = 3.32$; $t = -2.20$, $p = .04$).

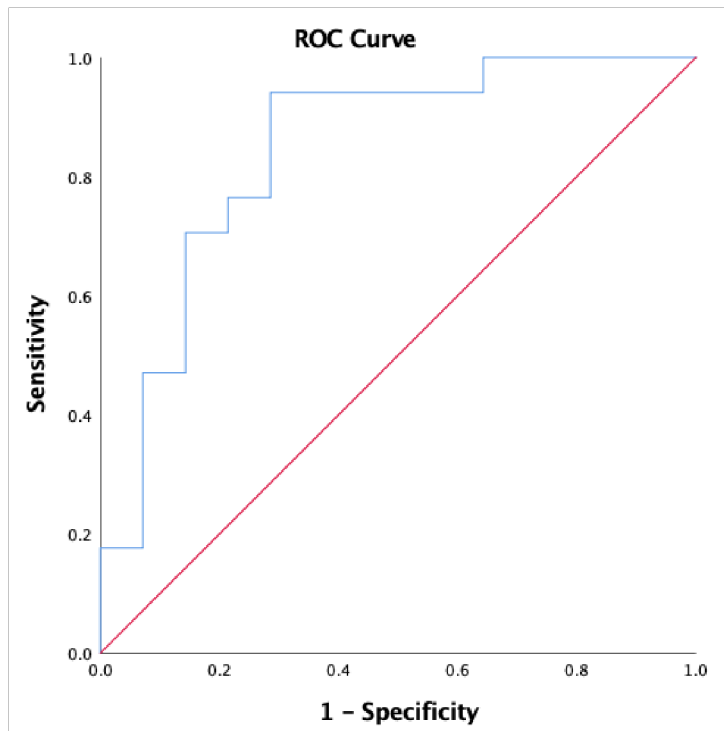
Research Question 2

The second research question examined the extent to which dynamic assessment of language in English and/or Spanish accurately identified school-age diverse children with LD (sensitivity) and without LD (specificity) when response to evidence-based English and/or Spanish language intervention over time was used to diagnose LD.

When the modifiability mean scores (total modifiability and the final modifiability scores) and the final intervention *NLM* score (posttest score) were entered into the logistic regression as predictors of the dynamic assessment, a combined probability variable was produced. We used that combined variables as the predictor in a receiving operator characteristic (ROC) area under the curve (AUC) analysis. The area under the curve was .85, with an optimal balance of sensitivity and specificity of 94% sensitivity and 71% specificity or 77% sensitivity and 79% specificity or 71% sensitivity and 86% specificity, see Figure 2.

Figure 2

Area Under the Curve (AUC) Output Indicating Optimal Balance of Sensitivity and Specificity



Discussion

The purpose of this study was to examine the accuracy of a dynamic assessment of language using modifiability ratings and gains in narrative language made through intervention over time as a primary indicator of a LD. An independent samples *t*-test indicated that between the first *NLM* and the last *NLM* administered during the intervention, there was no significant difference between the students with LD and the students with typical language development of language. However, results indicated that there was a significant difference in modifiability ratings between groups, with the children with LD having lower modifiability ratings than the students who did not have a LD. Also, the ROC analysis, which yielded an area under the curve, indicated that the sensitivity was 94% and specificity was 71%.

Narrative Language Measure Analysis

In this study, it was hypothesized that students with typical language development would make greater gains during intervention than the students with LD. The results of this study did not support this hypothesis. On the *NLM* both groups made similar amounts of gains over time through intervention, which suggests that intervention can possibly benefit children with and without LD. During the intervention, the interventionist utilized colored visual materials, active responding activities, and individualized interventionist support. These levels of support appear to be powerful enough to help students with LD make gains over time. It is possible that the students with LD could eventually reach and possibly surpass the students with typical language development if intervention were continued. Spencer and Slocum (2010) found that intervention can increase narrative language for students with LD.

It is possible that the children with LD made the same (or greater) gains than the children with typical language development because they had more to gain. Note how the children with LD had an initial mean *NLM* score of 10.00 whereas the children without LD had a mean score of 16.38. It is possible that there was a ceiling effect and that the TD children were already operating at their maximum capacity to tell stories (even though the test would allow for higher scores). If so, then the students with typical language would not make as strong a gain as a group of students that were not at this ceiling.

Response to intervention over time does not appear to validate results of the DYMOND. In fact, narrative-based language intervention improves narrative language in children who have a LD to the same or greater degree than TD students. Improvements in narrative skills is essential in children with LD (Spencer & Slocum, 2010). Because response to intervention in

these conditions does not differentiate students with LD from those without LD, other means of confirming LD would be necessary.

Modifiability Total Score

Even though the gain scores on the *NLM* were not significantly different, there was a significant difference between the mean total modifiability scores between children with and without LD. These lower modifiability scores for the children with LD indicate that the children with LD had greater frustration, less transfer of skills from one step to the next, paid less attention during the intervention, that it was harder for the examiner to teach the child, and that the child disrupted the intervention more often than children without a LD. Recall that the interventionists were blind as to whether the children did or did not have LD. Even so, their degree of effort and the degree to which the children struggled aligned with the DYMOND's classification of LD/no LD. Consistently across the dynamic assessment research, modifiability ratings have been the strongest indicators of LD. For example, in Peña et al. (2014) it was found that modifiability and posttest scores predicted language ability over all other variables. This was also found in Petersen et al. (2017). It appears that it is the effort required to help the students learn language more than the gains the students make that indicates a LD.

Sensitivity and Specificity

In order to determine the overall sensitivity and specificity of the DYMOND to response to evidence-based English language intervention over time, receiver operator characteristics (ROC) analyses were conducted. The AUC provides sensitivity and specificity for each possible cut point of the predictor measure. When attempting to identify the optimal balance, sensitivity and specificity were held at 70% or higher.

Although sensitivity was high (94%), specificity was only adequate (71%). However, with this small sample size of children, any misses in identifying children exaggerates the accuracy of the test. Consistent with other dynamic assessment studies, this study indicated that dynamic assessment of language can predict LD in bilingual, Spanish-English speaking children with high to moderate accuracy. For example, in a systematic review of dynamic assessments, Orellana et al. (2019) found high classification accuracy with dynamic assessment of language across six dynamic assessment studies.

Limitations and Future Research

More intervention sessions may have better established whether or not a child had a LD. The nine interventionists did not work with each student for the same amount of time, which could have affected their judgment of response to intervention. Intervention sessions varied by group based on teacher schedules and intervention sessions varied by individual student based on student attendance. There was a fairly small sample of students with and without typical language development in this study. Future research should include a greater number of participants. Even though the interventionists were trained over several hours on how to conduct the intervention, fidelity of intervention should be checked more carefully and quantified using a fidelity checklist in future research.

Conclusions

There are three major findings from this study. First it appears that the dynamic assessment results align with a student's response to intervention over time – not as measured by gain scores, but instead as measured using a modifiability rating. Second, students with LD when provided evidence-based narrative language intervention in a small group setting can make gains that are similar to those gains made by students without LD. In this study, most students only

received three or four intervention sessions. Based on their trajectory, it appears that if intervention were continued, they would eventually catch up to the TD peers since that were learning at a faster pace. Third, the dynamic assessment was able to identify nearly all the students with LD, even when disorder was established using a gold standard response to intervention process.

References

- Barragan, B., Castilla-Earls, A., Martinez-Nieto, L., Restrepo, M. A., & Gray, S. (2018). Performance of low-income dual language learners attending English-only schools on the clinical evaluation of language fundamentals-fourth edition, Spanish. *Language, Speech, and Hearing Services in Schools*, 49(2), 292-305.
- Bauman, K. (2017, August 28). *School enrollment of the Hispanic population: Two decades of growth*. United States Census Bureau, Census Blog. Census News.
https://www.census.gov/newsroom/blogs/randomsamplings/2017/08/school_enrollmentof.html
- Colby, S. L., & Ortman, J. M. (2015, March 3). *Projections of the size and composition of the U.S. population: 2014 to 2060*. United States Census Bureau.
<https://www.census.gov/library/publications/2015/demo/p25-1143.html>
- Dawson, J. I., Stout, C. E., & Eyer, J. A. (2003). *Structured Photographic Expressive Language Test (SPELT-3)*. Janelle Publications.
- Feuerstein, R., Rand, Y., & Hoffman, M. B. (1979). *The dynamic assessment of retarded performers: The Learning Potential Assessment Device Theory, instruments, and techniques*. University Park Press.
- Gardner, M. F. (1983). Expressive one-word picture vocabulary test. *Academic Therapy Publication*.
- Instituto Cervantes. (2018, November). *577 millones de personas hablan español, el 7,6 % de la población mundial*.
https://www.cervantes.es/sobre_instituto_cervantes/prensa/2018/noticias/np_presentacion-anuario.htm

- 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(1), 81-96.
<https://www.ncbi.nlm.nih.gov/pubmed/22052970>
- Kaufman, T. (1974). *Idiomas de mesoamerica*. Guatemala city: Seminario de integración social Guatemalteca.
- 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*, 33(3), 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(1), 44-55. [https://pubs.asha.org/doi/10.1044/0161-1461\(2003/005\)](https://pubs.asha.org/doi/10.1044/0161-1461(2003/005))
- National Center for Education Statistics. (2019, May). *English language learners in public schools*. https://nces.ed.gov/programs/coe/indicator_cgf.asp
- National Council on Disability. (2018). *IDEA Series: English learners and students from low income families*.
https://www.ncd.gov/sites/default/files/NCD_EnglishLanguageLearners_508.pdf
- Orellana C., Wanda, R., & Gillam, R.B. (2019). The use of dynamic assessment for the diagnosis of language disorder in bilingual children: A meta-analysis. *American Journal of Speech language Pathologists*. 28(3).1298-1317.
- 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 (Online)*, 57(6), 2208-2220.

<https://www.ncbi.nlm.nih.gov/pubmed/31194570>

- 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*(5), 1037-1057. 10.1044/1092-4388(2006/074
- Peña, E., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *Journal of Special Education, 26*(3), 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*(4), 983-998.
- Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48*(1), 3-21.
- Petersen, D. B. & Spencer, T. D. (2012, October 19). The narrative language measures: Tools for language screening, progress monitoring, and intervention planning perspective on language learning and education. *Perspectives on Language Learning and Education, 19*(4), 119-129. <https://doi.org/10.1044/lle19.4.119>
- Restrepo, M., & Kruth, K. (2000). Grammatical characteristics of a Spanish-English bilingual child with specific language impairment. *Communication Disorders Quarterly, 21*(2), 66-76. <https://journals.sagepub.com/doi/10.1177/152574010002100201>
- Restrepo, M., & 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 Version (CELF-4 Spanish)*. Pearson Education Inc.
- Sonnenschein, S., Metzger, Shari R., Dowling, R., Baker, L. (2017). The relative importance of English vs. Spanish language skills for low-income Latino English language learner's early language and literacy development. *Early Child Development Care*, 187, 727-743. <https://www.tandfonline.com/doi/abs/10.1080/03004430.2016.1219854?journalCode=gecd20>
- 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(1), 61-72. <https://pubs.asha.org/doi/10.1044/0161-1461%282006/007%29>
- Spencer, T. D. & Petersen, D. B. (2012). *Story Champs*. Language Dynamics Group.
- Spencer, T. D. & Slocum, T. A. (2010). The effect of a narrative intervention on story retelling and personal story generation skills of preschoolers with risk factors and narrative language delays. *Journal of Early Intervention*, 32(3), 178–199.
- Thorndike, R., Hagen, E., & Sattler, J. (1986). *Stanford-Binet intelligence scale* (4th ed.). Riverside
- 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(2), 142-154.
- U.S. Census Bureau. (2018, September 13). *Hispanic heritage month 2018*. <https://www.census.gov/newsroom/facts-for-features/2018/hispanic-heritage-month.html>

Cambridge, MA: Harvard University Press.

Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*.

Harvard University Press.

Wiley, T. G., Preyton, J.K., Christian, D., & Moore, S.C. (2014, January 6). *Handbook of heritage, community, and native American languages in the United States*. Routledge.

Williams, C. J., & McLeod, S. (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology*, 14(3), 292–305.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1993). *Preschool Language Scale-3, Spanish Edition (PLS-3 Spanish)*. The Psychological Corporation.

APPENDIX A

Annotated Bibliography

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*, 33(3), 119-128.

Objective: The purpose of this study was to probe the accuracy of the Dynamic Assessment of Instrument (Peña et al, 2001), dynamic assessment of narrative of language that Peña et al. (2006) and Peña et al. (2014) studied that was administered to a group of 3rd grade children from Samson Cree Nation Reserve in Alberta, Canada.

Methods: 17 children participated in this study, 5 of them were labeled as having a language impairment (LI) and 12 of them were classified as typical language development (TD). The five children were classified as having a language impairment based on the input of the special education teacher, the 3rd grade teachers, and the school's principal on each of the child's language status utilizing previous speech language pathology assessments, classroom performance and classroom observation. The administration of this Dynamic Assessment was equivalent to the dynamic assessment administration procedure as in Peña, (2001), Peña et al. (2006), and Peña et al., (2014). The administration of the entire dynamic assessment was finished in a period of 4 days. In this study they used the same wordless picture books as well as the scoring Likert scale to measure each student modifiability, responsiveness and narrative production as the

Dynamic Assessment and Intervention (DAI). The narrative transcripts of the Dynamic Assessment were scored by two examiners. The final scoring decisions of

pretest, posttest and modifiability were reached through consensus between the two examiners. However, interrater reliability on modifiability scoring and pre- and posttest scores were not reported.

Results: This study found that the Dynamic Assessment and Intervention was accurate for children in 3rd grade because it indicated classification accuracy in identifying children with language impairment. Although both groups had similar scores at the test phase, normal language students made greater improvements in targeted and non-targeted narrative elements. Modifiability ratings most accurately classified students, yielding 100% sensitivity and specificity was 92%. This study also shows that modifiability and responsiveness alone yielded 100% sensitivity, it was only 75% specificity.

Relevance to the current work: This study shows that the dynamic assessments are useful tests in identifying language differences ethnically diverse children.

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 (Online)*, 57(6), 2208-2220.

<https://www.ncbi.nlm.nih.gov/pubmed/31194570>

Objective: The purpose of this study was to evaluate Dynamic Assessment of English narration for children learning English as a 2nd language using the Dynamic Assessment and intervention (DAI) narrative learning task described by L. Miller , Guillam, and Peña (2001).

This study included four research questions:

1. What are the patterns of narrative learning from pretest to posttest based on language ability?
2. Are their differences in strategy use of language ability as indicated by observation of modifiability?
3. What combination of story and modifiability measures best differentiate children with language ability in matched and comparison samples?
4. Does the diagnostic accuracy of Dynamic Assessment differ between children without language impairment who were closely matched to the LI children and children without impairment who were not closely matched?

Method: 54 bilingual children participated in this study. 18 children with Language Impairment, 18 children with normal language development matched on age, sex, language experience and IQ (the Normal Language match group/NL-match), and another 18 children with normal language development match only in age and language experience (the NL-compare group). No second LI group. The Normal language match group was created by matching each of the children identified with Language Impairment to a TD child based on sex, age in months at time of initial testing, month of birth, IQ, and language experience which included percentage of English and Spanish input and output and age at which they had their first English exposure according to parent and teacher report. A second comparison control group (NL-compare) was created by matching a second NL child to each of the 18 children with LI using the age at which they had their first English exposure, age in months and percentage of English and Spanish input and output. For this comparison, a match on IQ or sex was not used as to improve the generalization of the cross-validation findings. Students were identified as

having language impairment by using the BESOS test results and by using narrative samples, and by evaluating teacher and parent responses about current language use at home and language proficiency.

They used the same Dynamic Assessment (DA) procedures as in Peña et al., (2006). The DA was conducted in three separate sessions over a 7-14-day period. The first session included the pretest and the first intervention session (MLE), the second session included the second intervention, and the third session was the posttest narrative. The two teaching (MLE) sessions were 30 minutes long and were completed in English. At the end of the first teaching session, the examiner completed a modifiability (responsivity) form. Each pretest and posttest was audio recorded, transcribed, and C-unit segmentation, then analyzed the samples using SALT (Total Number of Words (TNW), Number of Different Words (NDW), and Mean Length of Utterance in words (MLUw)) (this took some time...that is an important thing to bring up). They also analyzed story components, story ideas and language, and episode structure. These three-story analysis approaches were combined to yield a total DAI story score. Each item was rated using a 5-point scale based on the number components and ideas and language category. A 7point scale was also used to rate episode structure

Results: The results of this study show that the dynamic assessment yield to best classification in identifying children with language impairment. To identify the most parsimonious model, they use backward multiple-regression which resulted in, compliance, metacognition and task orientation. They also included three of the posttest scores, setting, knowledge of dialog and complexity of vocabulary and one of the five SALT story measures, ungrammaticality. The results of these 7 variables classified

88.9% sensitivity of children with language impairment and 88.9% specificity (children without language impairment). The cross-validated classification between the children with language impairment and the children in the NL-compare group shows 88.9% sensitivity and 72.2% specificity. In a second cross-validations, they found 100% sensitivity and 88.9% specificity with the language impair group and the NL-compare. They also cross-validated LI group with the NL-Match and that resulted in 100% sensitivity and 94.4% specificity.

Relevance to current work: The results of this study shows that the English narrative dynamic assessment is accurate in identifying bilingual Spanish-speaker children with language impairments. Additionally, when combining language sample Systematic Analysis of Language Transcript (SALT; Miller & iglesias, 2012), modifiability and posttest yielded the highest classification accuracy. Moreover, Interrater reliability scores were not reported in this study.

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(5), 1037-1057. 10.1044/1092-4388(2006/074

Objective: The purpose of this study was of two different experiments. The first explains whether parallel results of two wordless picture books yield to comparable measures without intervention sessions. The second, examines the extent to which children with language impairment performed differently than typically developing children on dynamic assessment of narrative language.

Methods: During the first experiment, 59 first and second grade diverse students from central Texas were asked to create a story from two wordless picture books. In this study, children were from different backgrounds including African American, European American and Latino American. The two groups were balanced for grade, 48% first graders and 52% second graders however, gender distribution was greater in girls over boys (64% girls vs 36% boys). Children in the typical developing group met at least three of the following criteria: (i) Teachers indicated no concerns regarding children's expressive and receptive language, and/or speech (ii) Parents indicated no concerns regarding children's expressive and receptive language, and/or speech via a questionnaire. (iii) classrooms observations using Patterson and Gilliam's (1995) of peer interaction indicated fewer than 15% syntactic, semantic, and/or pragmatic errors during a 10-minutes observation of play or group activity (iv) Children scored within one standard deviation of the mean in the Test of Language Development (TOLD-P-3) or the Comprehension of Spoken Language (CASL). The typical developing children group received the story, *Two Friends* ([L. Miller, 2000b](#)), followed by the story, *Bird and His Ring* ([L. Miller, 2000a](#)). While the second group of typical developing children received the story, *Bird and His Ring* ([L. Miller, 2000a](#)) first, followed by the story, *Two Friends* ([L. Miller, 2000b](#)).

For the second experiment, students were from different cultures including African American, European American and Latino American as stated by parents. Participants were divided into three different groups, a control group that consisted of 30 children, a typical developing group that consisted of 27 children and a language impaired group that consisted of 14 children. Children who were placed in the Language Impaired group met at least two of the following conditions: (i) Diagnosis of a language disorder by a certified Speech Language Pathology, (ii) parent's concern regarding

child's language expression and (iii) comprehension at school or at home, and (iv) performance less than or equal to 1.25 standard deviation below the mean on the Test of Language Development-Primary Third Edition (TOLD-P:3 (Newcomer & Hammill, 1997)). During pretest and posttest all children told a story based on two different books, *Two Friends* for the pretest, and *Bird and His Ring* for the posttest. Children in the typical developing group and language impairment group received two individualized 30-minute sessions focusing on narratives skills and strategies. At the end of the second intervention session, examiners evaluated how much support was required based on 5-point Likert scale. Child's responsivity was also evaluated on 5-point Likert Scale. A score of 5 meant high child responsivity and a score of 1 meant low child responsivity.

Results: The results for the first experiment showed that both books yielded comparable measures of children's narratives performance without intervention sessions. The outcome of the second study indicated that typical development group made a greater gain from pretest-posttest performance than children in the language impairment group and children in the control group. However, the gains that children in the language impairment group made after intervention were similar to those children in the control group that received no intervention. Children in the language impairment group made less gains from pretest to posttest performance than the children in the typical development group. Typical development children had higher modifiability scores than children in the language impairment group.

After posttest, the dynamic assessment narrative measures showed higher sensitivity values, 64%, than pretest story 26%. Posttest dynamic assessment narrative measures showed lower specificity, 83%, than pretest story, 88%. The most accurate

measure was the modifiability with 93% sensitivity and 82% specificity. The posttest scores and modifiability scores were combined to get the correct classification of number of different words, total number of words and the story components yielded 100% correct classifications meaning 100% accuracy in identifying children with language disorder.

Relevance to current work: The result of this study shows that the narrative dynamic administered had high classification accuracy in identifying language impairment in Englishspeaking children. Furthermore, the study also shows that sensitivity and specificity are highest when modifiability and post test scores are combined.

Peña, E., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *Journal of Special Education*, 26(3), 269–280.

Objective: Peña and Iglesias (1992) compared the efficacy of a dynamic assessment measure against a standardized assessment measure in identifying culturally diverse children with language disorder. Furthermore, the study also explored the mismatch between common linguistic tasks in standardized test measures and linguistic tasks common in Latino American and African American cultures.

Method: A total of 50 African American and Puerto Rican students from three Head Start classes in Northern Philadelphia participated in this study. All of the students were exposed to English and Spanish in the classroom. Two standardized test instruments were used: The Expressive One-Word Picture Vocabulary (EOWPVT; Gardner, 1979) and the Comprehension subtest of the Stanford-Binet Intelligence Scale (CSSB; Thorndike, Hage, & Sattler, 1986). The CSBS is a comprehension and description task while the EOWPVT elicits single-word labels. Researchers predicted that CSSB would

have better results with linguistics tasks more common to the student's home culture while the EOWPVT would be more foreign the children's home and cultural experiences. Those students who scored low on EOWPVT received mediation training (a dynamic assessment teaching phase). The mediation consisted of two 20-minute sessions that focused on improving vocabulary labeling abilities of the students. After each mediation session, the clinician scored each student based on their responsiveness, examiner effort, and transfer of skills to obtain an overall modifiability rating. After the two-mediation sessions, the students were assessed using the EOWPVT.

Results: The results of this study were reported using two different analyses. The first analysis indicated that both typically developing students and students with language disabilities scored equally low on the EOWPVT during the pretest. The students with language disorder scored lower in the CSSB. Second data showed that classification accuracy of the dynamic assessment was 92% (is this sensitivity?) of the language disordered cases. Finally, the typically developing children had higher modifiability scores as well as higher gains than the students with language impairment. This study shows that dynamic assessment is effective in determining language impairment in culturally diverse children. It also shows that pre-test standardized measures of assessment are less effective in differentiating between typically developing children and children with language impairment from different backgrounds.

Relevance to current work: This study demonstrated that a dynamic assessment of language is a more accurate assessment method than static measures. Norm-referenced measures are more biased in differentiating disorders from disability as many of the test items on these assessments are culturally foreign to diverse students. The dynamic

assessment of language can help us determine language differences from language disorder as well as academic needs. Additionally, this study also shows that 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(4), 983-998.

Objective: This article reports on the classification accuracy of an English narrative dynamic assessment for identifying Language Impairment (LI) in Spanish-English bilingual Kindergarten to third-grade students. The study used a more concise dynamic assessment with a real-time scoring procedure to indicate if LI could be identified in less time than conventional dynamic assessment measures as long as appropriate classification accuracy was maintained.

Methods: The study included 42 Hispanic children who were bilingual in both English and Spanish (10 with LI and 32 without LI) from a large urban school district in the mountain west. To evaluate their proficiency in both languages, language samples were analyzed using English and Spanish narrative retells. The students were classified as balanced bilingual, Spanish dominant or English dominant. In order for a student to have a diagnosis of language disorder, first a child had to have an IEP for language, second a bilingual SLP had to confirm this eligibility, third the student had to score below 1 SD below the mean in both languages on a narrative retell across at least one of the following: mean length of utterance, total number of words, and number of different words, lastly, oral or written confirmation of a language disorder from a parent or teacher.

The dynamic assessment was conducted within 2 days following the narrative retell. It consisted of two 25-minute test-teach-retest sessions. Each session consisted of a pretest narrative retell, a narrative retell teaching phase and posttest narrative retell. Both the pretest and posttest narrative retells and modifiability ratings were scored during the session. During the teaching phase, clinicians individually targeted story grammar and adverbial subordinate clauses. The pre- and post of the dynamic assessment were scored based on: (a) the nine story grammar elements (i.e., character, setting, problem, emotion, plan, attempt, consequence, ending and ending emotion) (b) occurrence of conjunctions (i.e., then, when, because, and after) and (c) complexity of episodic structure. The teaching phase targeted each of the elements used in the pre and posttests. After each teaching phase, the examiner scored the children using a modifiability rating scale used in previous dynamic assessment research.

Results: According to this article, the result of this study yielded high classification accuracy. The overall modifiability from both dynamic assessments' sessions yielded 100% specificity and 100% sensitivity. In addition, the modifiability score for one of the 25-minute sessions yielded to 100% sensitivity and 91% specificity.

Relevance to current study: The results of this study supports the assertion that the English dynamic assessments are accurate in identifying Spanish English children with language impairment. It also showed that there might be a possibility to shorten the dynamic assessment teaching phase so that it is more clinically useful. Furthermore, modifiability rating had a high interrater reliability and were more predictable 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*(1), 3-21.

Objective: The purpose of this study was to determine the extent to which a dynamic assessment of reading administered to kindergarten was predictive of reading difficulty at the end of first grade for bilingual Latino children. A second purpose of this study was to compare how accurate the dynamic assessment measure identified children with literacy deficits with the classification accuracy of a more traditional, static measure of reading ability.

Method: 63 Latino bilingual kindergarten children participated in this study.

These children were identified at risk for language impairment during pre-kindergarten screening. The children attended general education classrooms in which English was the primary language of instruction. All of the participants were English language learners before entering kindergarten and all of them had lived in the US for at least one year. The Bilingual English-Spanish Assessment (BESA; Peña, Gutierrez-Clellen, Iglesias, Goldstein, and Bedore, 2014) was administered to each of the participants. Children who scored below the 30th percentile were considered at risk for language impairment. The dynamic assessment of reading consisted of a pretest that assessed their ability to read nonsense words, a teaching phase, and a posttest using the same words. In the teaching phase, children were taught reading strategies using nonsense words. The children were asked to recode words used in the pretest phase by using an onseprime, analogous strategy in conjunction with whole word recognition. During the posttest phase the children were asked to recode the same nonsense words used in the pretest and teaching phase, displayed in a different order. Each participant was scored on gains from pretest to

posttest. The participant's ability to read words was also assessed using reading strategy analysis. Each student was rated on a 3-point response to instruction scale based on each examiner's perception of how difficult it was for the participant to respond to the instruction, ranging from easy to difficult. The reading strategy score was combined with the response to instruction scores to create a dynamic assessment modifiability score. Interrater agreement on the scoring of total number of correct sounds, words, and response to instructions was 97%. Interrater agreement on the scoring reading strategy was 98%. Additionally, the participant's reading abilities were assessed by using static subsets from the Dynamic Indicators of Basic Literacy Skills (DIBELS) standardized assessment.

Results: the strongest predictor of first grade reading was the dynamic assessment modifiability scored. The residuum gain score was also predictive of the first-grade reading measures. However, the dynamic assessment sound gain score was not predictive of first grade reading. The modifiability score of the dynamic assessment of reading had the highest validity in predicting which bilingual Latino children would be at risk for reading difficulty at the end of first grade. The modifiability score yielded 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 non-word fluency scores in first grade. In contrast, the static kindergarten DIBELS measure used to assess the children's literacy resulted in high overclassification of students as at risk for reading difficulty.

Relevance to work: This study shows that dynamic assessment measures have a higher classification accuracy than static measures for bilingual students. In this study

static measures had over-classification of bilingual student. In contrast, dynamic assessment modifiability scores had the highest sensitivity and specificity. Modifiability was the strongest predictor of reading ability. Were dynamic assessment posttest scores also predictive of reading?

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*(2), 142-154.

Objective: the purpose of this study was to examine if the dynamic assessment intervention was a more culturally appropriate measure of language ability than standardized tests for 23 Arapahoe and Shoshone kindergarten children.

Method: Twenty-three kindergarten children from an elementary school on the Wind River reservation in Wyoming participated in this study. 15 of the twenty-three children were considered as a stronger language-learner group and the rest, 8 children, were considered a weaker language learner group. English was the primary language spoken by these students. Even though the primary language for these students was English, they had some exposure to the other two languages spoken at home or school. A test-teach-test dynamic assessment was administered in a period of 3-weeks. The testing took approximately 20-minutes each time for each child. The two mediation (teaching) sessions lasted 30-minutes each and the children were seen in pairs. Each mediation session entailed teaching the students vocabulary categorizing skills by learning to group similar words under a unifying category. After each mediation, the examiners scored each student's positive learning behaviors and positive responses to instruction by using a 5-point Likert Scale. A standardized NRTs, Everyday Themes (ASSETS; Barrett, 1988),

was used to assess semantic skills as pre and post measures to investigate the effect of the mediation sessions. This assessment was administered once 1 to 5 days prior to the dynamic assessment mediation sessions and once 1 to 5 days after the administration of the final mediation session of the dynamic assessment. Inter-rater reliability in scoring modifiability, learning strategies and responsiveness, was 94%. Inter-rater reliability in scoring the ASSET was 96%.

Results: The results of this study showed that the modifiability scores and the post-test scores of the ASSET were higher for the stronger language-learner group than the scores of the weaker language-learner group. This study also showed that the student's positive response to instruction was a greater predictor of the difficulty of learning language than the student's positive learning behaviors measured during the teaching phase. The specificity and sensitivity of the dynamic assessment was not reported.

Relevance to current work: this study shows that a dynamic assessment of vocabulary can be used to differentiate language differences from disorders in culturally diverse students. Although the classification accuracy of this assessment is unknown, this study supports the assertion that students with typical language learn language with less difficulty compared to students with language disorders.

APPENDIX B

CUBED Narrative Language Measures: Listening (NLM: Listening)

NLM LISTENING		Kindergarten Benchmark: STORY 1		WINTER								
Child/ID _____ Audio File _____ Examiner _____ Date _____												
LISTENING RETELL	Examiner says, "I'm going to tell you a story. Please listen carefully. When I'm done, you are going to tell me the same story. Are you ready?" Examiner reads the story word for word at a moderate pace with normal inflection.											
	<p> Yesterday, Holly and her friend clambered onto the bus. They quickly went up the steps because the bus was about to leave. Holly's friend sat in the seat that was by the window. But Holly didn't want her to sit there because it was her favorite seat. She was mad. Her friend was in her usual window seat. Holly decided to politely talk to her. Then she said, "Excuse me. Will you please move? I typically sit there." Then her friend said, "Okay. No problem. You can sit by the window." After her friend moved, Holly sat adjacent to the window. When Holly sat down, she was happy because she could see out the window. </p>											
Examiner says, "Thanks for listening. Now you tell me that story." After student appears to be done, examiner says, "Are you finished?" Prompts (up to 3x), "It's OK. Just do your best." and/or "I can't help, but you can just tell the parts you remember."												
LISTENING RETELL	STORY GRAMMAR (SG) 2 POINTS		1 POINT		0		LANGUAGE COMPLEXITY (LC)		EPISODE (E)			
							Word #Times Used		(from green 2 point SG)			
	Character	Holly / any name	2	a girl / the girl	1	0	because	1 2 3	P+A P+C A+C	2		
	Setting	climbed into the bus / climbed into the bus at school	2	at school / in the bus / climbed	1	0	when	1 2 3	P+C+E P+A+E	3		
	Problem	friend was in her seat	2 [P]	couldn't sit	1	0	after	1 2 3	P+A+C	4		
	Feeling	sad / mad / angry	2	didn't like it / cried	1	0	LC SUBTOTAL		P+A+C+E	5		
	Plan		-	planned / decided	1	0	OTHER TARGETS		E SUBTOTAL			
	Attempt	asked her friend to move	2 [A]	asked for help	1	0	Target #Times Used					
	Consequence	said "Sure, I'll move."	2 [C]	she helped	1	0	Then	✓				
	Ending	got to sit by window / could see out the window	2 [E]	did it	1	0	Modifiers	✓				
End Feeling	happy / excited	2	felt better / liked it	1	0							
SG SUBTOTAL								RETELL SCORE (SG+LC+E)				
COMPREHENSION QUESTIONS	STORY QUESTIONS (SQ) 1x		VOCABULARY QUESTIONS (VQ) 1x		3 = clear 2 = unclear 1 = correct 0 = incorrect		3 = clear 2 = unclear 1 = correct 0 = incorrect		3 = clear 2 = unclear 1 = correct 0 = incorrect			
	Who was this story about?		2	1	0	QA: They clambered onto the bus because it was about to leave. What does clamber mean?"		3	2			
	Where was Holly in the beginning of the story?		2	1	0	QB: Does clamber mean to fall down or to climb?"		1	0			
	Why was Holly mad?		2	1	0	QA: Holly's friend was in Holly's seat. Holly typically sits there. What does typically mean?"		3	2			
	What did she do to fix the problem?		2	1	0	QB: Does typically mean usually or never?"		1	0			
	How did the story end?		2	1	0	QA: Holly sat adjacent to the window. She looked out the window. What does adjacent mean?"		3	2			
	What will Holly do the next time someone is in her favorite seat?		2	1	0	QB: Does adjacent mean next to or backwards?"		1	0			
	STORY QUESTIONS TOTAL (SQ)						VOCABULARY TOTAL (VQ)					
PERSONAL GENERATION		(Turn on audio recorder). Examiner says, "In this story, someone was sitting in Holly's seat. Tell me a story about a time when someone was in your seat." If the student doesn't tell a story, encourage the student (up to 3x) to produce a thematically related story. Score the story using the NLM Flow Chart (see Examiner's Manual for details).										

APPENDIX C

Small Group Narrative Intervention Fidelity Checklist

SMALL
GROUP
MASTER LESSON PLAN
26


TARGET

Enhanced Story Structure


Consider ADD ON lessons 58-63

MATERIALS


- ✓ Choose any **CLASSIC** or **BLITZ Level B** story from story book
- ✓ **Illustrations**
 - If using illustration cards, select cards from corresponding story (for **BLITZ** stories, use only **cards 1, 2, 3, 7 and 8**)
 - If using digital presentation, click on the purple **Level B** button and select the corresponding story
- ✓ **Story Grammar Icons** (icons are included in the digital presentation)




character




setting




problem




feeling



action



ending



end feeling
- ✓ Choose a **Story Game**
 - Each student should have 1 cube, 1 bingo card, **OR** 7 sticks (game materials are not needed to play Story Gestures)

1 – Model Story

- ☐ Display 5 illustrations
- ☐ Read the story
- ☐ Place Story Grammar icons on or near illustrations
- ☐ As needed: Name the Story Grammar parts
- ☐ As needed: Students name the Story Grammar parts

2 – Team Retell

- ☐ Leave illustrations on table
- ☐ Pick up icons and give each student 1-2 icons; keep one for yourself if necessary
- ☐ Starting with the person who has the Character icon and moving through the parts in order, each person retells the part of the story
- ☐ Students place icons on or near illustrations
- ☐ Summarize the story quickly and ensure that all parts are included

3 – Individual Retell 1

- ☐ Leave illustrations and icons on table
- ☐ Select one student to retell entire story
- ☐ Help the student retell all parts of the story
- ☐ Everyone, but the storyteller, plays a Story Game
- ☐ Summarize the story quickly and ensure that all parts are included

4 – Individual Retell 2

- ☐ Remove illustrations and leave icons on table
- ☐ Select one student to retell entire story
- ☐ Help the student retell all parts of the story
- ☐ Everyone, but the storyteller, plays a Story Game
- ☐ Summarize the story quickly and ensure that all parts are included

5 – Individual Personal Story 1

- ☐ Leave icons on table
- ☐ Select one student to tell a personal story
- ☐ Say, "Has something like that every happened to you?"
- ☐ Help the student generate all parts of the student's personal story
- ☐ Everyone, but the storyteller, plays a Story Game
- ☐ Summarize the student's story

6 – Individual Personal Story 2

(skip if fewer than 4 students)

- ☐ Remove icons from table
- ☐ Select one student to tell a personal story
- ☐ Say, "Has something like that every happened to you?"
- ☐ Help the student generate all parts of the student's personal story
- ☐ Everyone, but the storyteller, plays a Story Game
- ☐ Summarize the student's story

REMEMBER!

- ✓ Assign students to steps 3-6 so the order in which they retell and tell stories changes frequently
- ✓ Use **2-Step Prompting** to help students
 - 1) Ask a question
 - 2) Model what the student should say
- ✓ Make corrections immediately
- ✓ Differentiate targets for each student

APPENDIX D

IRB Approval Form



INSTITUTIONAL REVIEW BOARD
FOR HUMAN SUBJECTS

Memorandum

To: Professor Douglas Petersen
Department: COMD
College: EDUC
From: Sandee Aina, MPA, IRB Administrator
Bob Ridge, PhD, IRB Chair
IRB#: X17484

Title: *"The Classification Accuracy of an English and Spanish Narrative Dynamic Assessment for Diverse School-Age Students"*

Brigham Young University's IRB has renewed its approval of the research study referenced in the subject heading. The approval period is through **March 7, 2020**. All conditions for continued approval during the prior approval period remain in effect. These include, but are not necessarily limited to the following requirements:

1. A copy of the consent forms are attached to this email. No other forms should be used. Each research subject must sign the form prior to initiation of any protocol procedures. In addition, each subject must be given a copy of the signed consent form.
2. Any modifications to the approved protocol must be submitted, reviewed, and approved by the IRB before modifications are incorporated in the study.
3. In addition, serious adverse events must be reported to the IRB immediately, with a written report by the PI within 24 hours of the PI's becoming aware of the event. Serious adverse events are (1) death of a research participant; or (2) serious injury to a research participant.
4. All other non-serious unanticipated problems should be reported to the IRB within 2 weeks of the first awareness of the problem by the PI. Prompt reporting is important, as unanticipated problems often require some modification of study procedures, protocols, and/or informed consent processes. Such modifications require the review and approval of the IRB.

IRB Secretary
A 285 ASB
Brigham Young University
(801)422-3606