2019-10-01

The Cross-Validation of the Classification Accuracy of a Dynamic Assessment of Narrative Language for School-Age Children with and Without Language Disorder

Kallie Dawn Clark
Brigham Young University

The Cross-Validation of the Classification Accuracy of a Dynamic Assessment of Narrative Language for School-Age Children with and Without Language Disorder

Kallie Dawn Clark

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
Tyson G. Harmon
Ryan O. Kellems

Department of Communication Disorders

Brigham Young University

ABSTRACT

The Cross-Validation of the Classification Accuracy of a Dynamic Assessment of Narrative Language for School-Age Children with and Without Language Disorder

Kallie Dawn Clark
Department of Communication Disorders, BYU
Master of Science

Purpose: This study examined how well a dynamic assessment of narrative language accurately identified kindergarten through sixth grade students with and without language disorder. Method: The participants included 110 school-age children from Utah and Colorado who were administered a narrative-based dynamic assessment of language that entailed a pretest, a teaching phase, an examiner rating of the child's ability to learn language (modifiability), and a posttest. Results: The dynamic assessment investigated in this study demonstrated good to excellent levels of sensitivity and specificity. The results of this study also determined that, in concurrence with previous dynamic assessment research, posttest and modifiability scores were most predictive of language ability. Conclusion: The results of this study indicate that the Dynamic Measure of Oral Narrative Discourse (DYMOND) may be a valid and accurate tool when identifying language disorders in school-age populations.

Keywords: oral language, narrative, dynamic assessment

ACKNOWLEDGMENTS

The author would like to thank the following committee members for their meaningful contribution: Douglas B. Petersen, Tyson G. Harmon, and Ryan O. Kellems. Their expertise, guidance, and experience were vital to the success of this project. The author also thanks the members of the DYMOND team for their time and effort in collecting the data used in this project.

TABLE OF CONTENTS

TITLE PAGE i

ABSTRACT ii

ACKNOWLEDGMENTS iii

TABLE OF CONTENTS iv

LIST OF TABLES vi

DESCRIPTION OF THESIS STRUCTURE vii

Introduction 1

Norm-Referenced Tests and Culturally and Linguistically Diverse Children 3

Dynamic Assessment 6

Dynamic Assessments of Narrative Language 9

Method 12

Participants 12

A-Priori Language Disorder 12

Measures 13

CUBED: Narrative Language Measures (NLM). 14

Non-word repetition task. 14

DYMOND 15

Dynamic assessment pretest. 15

Dynamic assessment teaching phase. 15

Dynamic assessment modifiability 16

Dynamic assessment posttest. 17

Test Administration: Fidelity and Inter-Rater Reliability.. 17

Fidelity. 17

Inter-rater reliability. 18

Results.. 19

Discussion 21

Clinical Implications 22

Need for valid language assessments. 22

Clinician concerns with dynamic assessment efficiency. 23

Clinician concerns with reliability. 23

Role of dynamic assessment in informing treatment. 24

Study Limitations and Future Research 24

References 26

APPENDIX A: Annotated Bibliography 31

APPENDIX B: IRB Approval 46

LIST OF TABLES

Table 1 *Demographic Information* 13

Table 2 *Logistic Regression and ROC Analyses for Narrative Dynamic Assessment Predictor Variables*.. 21

DESCRIPTION OF THESIS STRUCTURE

To adhere to traditional thesis requirements and journal publication formats, this thesis, *The Cross-Validation of the Classification Accuracy of a Dynamic Assessment of Narrative Language for School-Age Children with and Without Language Disorder,* is written in a hybrid format. The initial pages of the thesis adhere to university requirements while the thesis report is presented in journal article format. An annotated bibliography is included in Appendix A. Appendix B includes approval documentation from the Institutional Review Board (IRB).

Introduction

Currently, norm-referenced tests (NRTs) are most commonly used to identify school-age children with language disorders. Many of the available NRTs lack an adequate level of sensitivity and specificity for identifying language disorders in school-age children (Spaulding, Plante, & Farinella, 2006). Spaulding et al. (2006) reviewed and analyzed the sensitivity and specificity of 43 norm-referenced language tests. In the context of language testing, sensitivity is the ability of a test to correctly identify students who have a language disorder. Specificity is the ability of a test to correctly identify students without a language disorder. Spaulding et al. noted that diagnostic assessments should yield sensitivity and specificity levels that are at or above 80% in order to have adequate evidence of classification validity. Of the 43 tests analyzed, they found only 5 that yielded an acceptable level of sensitivity and specificity (≥80%). When assessments lack adequate sensitivity and specificity, the number of false positives (Type 1 errors) and false negatives (Type 2 errors) increases. In short, the chance of misidentifying children increases when test measures lack adequate sensitivity and specificity.

Denman et al. (2017) identified which norm-referenced language assessments had the best evidence for clinical use. The researchers originally found 76 language assessments to be evaluated. However, the researchers imposed certain exclusionary criteria to improve the quality and detail of their study. They excluded (a) assessments (n = 61) that were not published in the 20 years prior to the study, (b) screening assessment tools, (c) outdated versions of assessments (i.e., included only the most current editions of each language assessment), and (d) assessments of academic achievement. The researchers also intentionally selected assessments that were considered to be comprehensive (i.e., assessments which had both expressive and receptive subtests or outcomes). After imposing these exclusionary criteria, Denman et al. (2017) were left with 15 language assessments. The researchers reviewed the psychometric quality of each of the 15 language assessments by rating each assessment based on internal consistency, reliability, error measurement, content validity, structural validity and hypothesis testing (i.e., reliability, validity responsiveness and interpretability). The researchers concluded that, of the 15 language assessments that they analyzed, all of the assessments lacked evidence of psychometric quality.

In 2010, Friberg conducted a similar study. Only tests which were found to have adequate sensitivity and specificity (>80%) were included in the study. Friberg also utilized the criteria from Spaulding et al. (2006) to identify which language assessments to evaluate for psychometric quality. After imposing such criteria, nine language assessments were identified for review and evaluation. The Clinical Evaluation of Language Fundamentals-4 (CELF-4), Clinical Evaluation of Language Fundamentals-Preschool Edition 2 (CELF-P2), Preschool Language Scales-4 (PLS-4), Structured Photographic Expressive Language Test-3 (SPELT-3), Structured Photographic Expressive Language Test-Preschool Edition 2 (SPELT-P2), Test for Examining Expressive Morphology (TEEM), Test of Early Grammatical Impairment (TEGI), Test of Language Competence-Expanded Edition (TLC-E), and Test of Narrative Language (TNL) were chosen for further psychometric evaluation. Psychometric criteria (11) were applied to each of the selected tests in order to determine each test's overall level of psychometric validity. The criteria were (a) the purpose of the test was identified, (b) the test qualifications were explicitly stated, (c) the testing procedures were adequately explained, (d) an adequate standardization sample size (>100) was noted, (e) a clearly defined standardization sample including information related to geographic representation, socioeconomic status, parent education, gender distribution, ethnic background, presence/absence of impairment and age was included, (f) evidence of item analysis, (g) measures of central tendency were reported, (h) concurrent validity was documented, (i) predictive validity was documented, (j) test/re-test reliability was reported, and (k) inter-rater reliability was reported. Friberg found that, of the 9 assessments that were evaluated, no assessment tool met all 11 psychometric criteria. However, all of the assessments met 8-10 of the criteria. Although most assessments reflected an acceptable level of each criteria, it is important to note that 78% of the evaluated language tests did not meet predictive validity criteria and 44% did not meet the test-retest criteria. These results indicate that there are still improvements to be made in the validity and reliability of language assessment.

Norm-Referenced Tests and Culturally and Linguistically Diverse Children

The population in the United States is increasing in diversity, and the Hispanic, Spanish-speaking population is responsible for the majority of this increase (Colby & Ortman, 2015; U.S. Census Bureau, 2019). Norm-referenced tests are often biased against children who are culturally and linguistically diverse (Laing & Kamhi, 2003; Petersen & Gillam, 2015). There is an increasing need to develop valid and reliable language assessments, for all children, including school-age children who are culturally and linguistically diverse.

Although many NRTs fail to successfully identify school-age children with and without language disorders, NRTs are particularly poor at identifying language disorders in children who are culturally and linguistically diverse. This poor classification of culturally and linguistically diverse students can be attributed to a variety of factors, including cultural and linguistic bias of test items, materials, and procedures. Laing and Kamhi (2003) discuss three problems that NRTs have when administered to culturally and linguistically diverse children. These problems include content bias, linguistic bias, and disproportionate representation of culturally and linguistically diverse populations in normative samples. Content bias occurs when a test and its methods, procedures, and/or test items assume that all children who are administered the test have been exposed to the same concepts, vocabulary, and life experiences. As most test items and procedures are derived from white-middle class culture, bias toward those who are culturally diverse can occur. Thus, children who are culturally diverse may experience content bias from NRTs because of differences in several factors, including life experiences, culture, or exposure to literacy experiences. Linguistic bias occurs when there is a mismatch between the language or dialect that the child speaks and the language or dialect that the examiner and/or test expects from the child. These linguistic biases may lead to misclassification. Disproportionate representation of culturally and linguistically diverse populations in the normative samples of many NRTs is also responsible for much of the bias against culturally and linguistically diverse children. Historically, NRTs have excluded culturally and linguistically diverse children from normative samples. In an attempt to better represent culturally and linguistically diverse populations in the United States, many test makers have recently made an effort to include diverse populations in their normative samples. Despite these attempts, the normative samples of many NRTs still underrepresent culturally and linguistically diverse populations (Laing & Kahmi, 2003). Laing and Kahmi suggest that these biases may only be eliminated by creating test measures which are specifically designed for culturally and linguistically diverse populations.

It is evident that the majority of English NRTs of language are biased and unsuited to assess culturally and linguistically diverse students. However, using available NRTs in a student's native language has also failed to yield adequate and acceptable levels of sensitivity and specificity. This is the case with the Spanish Preschool Language Scale (SPLS-3; Zimmerman, Steiner, & Pond, 1993) and the Clinical Evaluation of Language Fundamentals

Fourth Edition-Spanish Version (CELF-4S; Semel, Wiig, & Secord, 2006), two commonly used NRTs. Restrepo and Silverman (2001) evaluated the validity of the Spanish edition of the Preschool Language Scale when administered to bilingual Spanish-speaking children with and without a language disorder. They found that the mean scores of the children in the study were significantly lower than the means of the Hispanic normative group in the SPLS-3. Furthermore, 51% of the children in the study performed at least one standard deviation below the mean of the SPLS-3 normative Hispanic group. This means that the children in the Restrepo and Silverman study, despite the fact that the majority had typical language development, scored on the low end of SPLS-3 distribution. Furthermore, they found a lack of construct and content validity. In their evaluation, many test items were deemed culturally inappropriate (e.g., stamps, money, wagon) and developmental expectations (e.g., vocabulary, grammatical markers) for Spanish language were incorrectly assumed to match those of English language development. Furthermore, the SPLS-3 lacked criterion validity when compared to other criterion-referenced measures, including language samples and parent/teacher concern.

Barragan, Castilla-Earls, Martinez-Nieto, Restrepo, and Gray (2018) conducted a study on the Clinical Evaluation of Language Fundamentals Fourth Edition-Spanish Version (CELF-4S; Semel et al., 2006), the most commonly used Spanish NRT. In their study, they found that the CELF-4S lacked adequate classification accuracy when used to identify language disorder in bilingual Spanish-speaking elementary school-age children from low socioeconomic backgrounds. The CELF-4S manual suggests using a cut off standard score of 85, one standard deviation below the mean. When using this score as the cut off, the researchers found that the CELF-4S yielded a sensitivity level of 94% and a specificity level of 65%. The findings of Restrepo and Silverman (2001) and Barragan et al. (2018) indicate that NRTs that are designed for Spanish-speaking populations may lack adequate levels of sensitivity and specificity. Thus, changing the language of the test and of the administration does not necessarily eliminate problems and biases associated with NRTs. Alternative measures need to be explored to mitigate current problems associated with NRTs and culturally and linguistically diverse populations.

Dynamic Assessment

Dynamic assessment is a promising alternative to NRTs. Dynamic assessment has been shown to mitigate bias in identifying school-age children with language disorders because, unlike NRTs, dynamic assessment does not assess a child's static, current knowledge. Rather, dynamic assessment measures a child's ability to learn and respond to evidence-based instruction. Dynamic assessment often uses a test-teach-retest model (Peña & Iglesias, 1992). The dynamic assessment begins with a pretest which assesses a child's current ability. Unlike NRTs, a dynamic assessment employs a teaching phase following the initial test. In the teaching phase, the child receives interactive and individualized instruction on the target behavior. During the teaching phase, the examiner provides the necessary support that the child needs to be successful and fades the support as the child learns throughout the teaching phase. During the teaching phase, the examiner also pays attention to how difficult it is for a child to learn and respond to direct instruction. Following the teaching phase, the examiner reflects on the teaching experience and completes a modifiability rating. Modifiability is a combination of how difficult it is for the examiner to teach the child and how difficult it is for the child to respond to intervention. Following the teaching phase, a posttest is administered which is similar in complexity and structure to the pretest.

Dynamic assessment is influenced by principles of Vygotsky's Zone of Proximal Development (ZPD; Vygotsky, 1978) and Feuerstein's Mediated Learning Experiences

(Feuerstein, Rand, & Hoffman, 1979). A child’s ZPD falls between what a child can do independently and what a child can do only with support. Vygotsky’s theory suggests that learning occurs when a child is instructed in his/her ZPD. Vygotsky’s theory when applied to dynamic assessment indicates that children should make gains from pretest to posttest as a result of direct instruction within their ZPD. When applied to dynamic assessment, this learning is often measured through the administration of a posttest.

Feuerstein’s Mediated Learning Experiences (MLE; Feuerstein et al., 1979) involve intentional instruction with the goal of facilitating independent learning in children. Feuerstein’s approach focuses on the child’s behavior (response to instruction) during the MLEs, as opposed to Vygotsky’s (1978) focus on learning outcomes measured through a posttest. When applied to dynamic assessment, a child’s attention, response to prompts, metacognition (awareness of errors), disruptive behavior, transfer of strategies within and between tasks, motivation, and level of frustration can be observed (Peña et al., 2006; Peña, Gillam, & Bedore, 2014; Petersen, Chanthongthip, Ukrainetz, Spencer, & Steeve, 2017).

MLEs require the examiner to attend to three variables: intentionality and reciprocity, meaning, and transcendence. Intentionality requires the examiner to be alert, vigilant, and animated while teaching explicitly. Reciprocity is engaging the examinee in the learning experience by responding to the individual and making the assessment an interaction. Transcendence is the development, use, and generalization of metalinguistic skills (e.g., an individual may identify and describe story grammar elements and then generalize that knowledge independently when telling a story). Meaning is explaining the “why” behind the assessment and teaching experience (e.g., “I need you to tell me the character so that I know who you are talking about”). Transcendence is more than learning a task—it is being able to transfer the behavior changes to other contexts (Feuerstein et al., 1979). This involves a shift from extrinsic to intrinsic motivation. In other words, the individual is motivated during the learning experience because they understand why they are doing it. This intrinsic motivation is fostered by providing encouragement and ensuring success through frequent, explicit feedback.

While Vygotsky's (1978) approach focuses on gains and changes in task performance, Feuerstein's approach focuses on the assessment of changes in student behavior during the learning experience. Both Vygotsky's and Feuerstein's approaches are taken into account when assessing a child's response to intervention using a test-teach-retest dynamic assessment. Pretest and posttest scores provide measures of gains (Vygotskian), and modifiability ratings capture the child's behavior and behavioral change during the teaching phase (Feuerstein). When a child begins to produce the behaviors taught by the examiner with less support, it indicates that learning, or transcendence, has occurred and the child has responded to the intervention (Peña et al., 2006).

The modifiability outcomes from dynamic assessments are unique. By measuring learning ability rather than comparing static performance to peers, much of the bias against culturally and linguistically diverse children is mitigated. Because of the potential benefit of dynamic assessment, several studies have investigated the potential of dynamic assessments of language to identify language disorders in children. Peña and Iglesias (1992), Kapantzoglu, Restrepo, and Thompson (2012) and Ukrainetz, Cooney, Dyer, Kysar, and Harris (2000) investigated the classification accuracy of dynamic assessments of vocabulary. Narrative language, however, has been the primary focus of dynamic assessment of language research because it is a highly effective and functional measure of language ability.

Dynamic Assessments of Narrative Language

There is an emerging body of evidence supporting the use of English dynamic assessments of language, particularly with culturally and linguistically diverse children, when narration (storytelling) is used as the assessment medium (e.g., Hasson & Joffe, 2007; Peña, Iglesias, & Lidz, 2001; Patterson, Rodríguez, & Dale, 2013; Peña & Iglesias, 1992). Research indicates that a dynamic assessment of narrative language can be accurate in identifying diverse children with and without language disorders. There are many benefits to using narratives to assess language ability. Rather than measuring fractionalized, individual language skills, narratives allow for the measurement of integrated, academic language in a functional and meaningful communicative context (Ukrainetz et al., 2000; Westby, 1985). Narratives involve the use of story grammar elements, which are important components that make up a story. Story grammar elements include character, setting, problem, action, consequence and resolution, in addition to feelings. Narratives also involve the use of complex, academic language to provide context for a story. Language complexity elements, such as adverbial and adjectival subordinate clauses, are important in clearly and efficiently providing details for the narrative.

Miller, Gillam, and Peña (2001) published a dynamic assessment with the intent of creating a culturally sensitive tool to assess language learning ability. This dynamic assessment applied the principles of mediated learning experiences. In 2006, Peña et al. used this dynamic assessment to identify which dynamic assessment test variables were most predictive of language disorder. Peña et al. administered this dynamic assessment to 71 diverse first and second graders, including students who were Hispanic/Latin-American, African American, and European American. The dynamic assessment consisted of three phases: a pretest, a teaching phase, and a posttest. The pretest required a child to create a story using a wordless picture book. The examiner then analyzed the child’s narrative language sample. The teaching phase utilized two, 30-minute mediated learning sessions that focused on storytelling. The posttest required the child to produce a new story using another wordless picture book. The researchers found that modifiability and posttest scores provided 100% sensitivity and 100% specificity.

In 2009, Kramer, Mallett, Schneider, and Hayward conducted a replication study of Peña et al. (2006). The dynamic assessment was administered to 17 third-grade students who lived in First Nations community. A discriminant analysis revealed high levels of sensitivity and specificity, similar to the results found in Peña et al. (2006). Peña et al. (2014) investigated the same English dynamic assessment of narrative language used in Peña et al. (2006), Peña et al. (2014), and Kramer et al. (2009). The dynamic assessment, which took place over the span of three days, was administered to 54 bilingual children. Of the 54 children, 18 spoke Spanish and English and had language disorders, and 18 were age-, sex-, IQ-, and language experience-matched. An additional 18 students were age- and language-experience matched comparison participants. The researchers found that, like in previous research, the modifiability score and posttest scores best predicted language disorders. They found sensitivity ranged from 89.9% to 100% and specificity ranged from 72.2% to 94.4%, depending on the groups of typically developing children that they included in the analysis. In general, their research indicated that an English dynamic assessment of language can identify diverse children with language disorders with adequate sensitivity and specificity.

Most recently, Petersen et al. (2017) conducted a study in which they investigated the classification accuracy of a brief (i.e., single 10-20 minute session) dynamic assessment called the Dynamic Measure of Oral Narrative Discourse (DYMOND). The assessment was administered to 42 Spanish-English bilingual children in kindergarten to third grade. Of the 42 participants, 10 had language disorder and 32 were typically developing. The Petersen et al. dynamic assessment consisted of a pretest, a teaching phase, and a posttest. In the pretest, the child was told a story and asked to retell that story out loud. The examiner scored the child's retell in real time, marking whether or not the child had included a variety of story grammar elements and language complexity targets (e.g., because, when, after). Following the pretest, the examiner used a structured intervention approach to teach the child to include various story grammar elements and targets which reflect linguistic complexity. Immediately following the teaching phase, the examiner briefly rated how difficult it was to teach the child and how difficult it was for the child to learn during the teaching phase. Lastly, the child was given a posttest, which followed the same procedures as the pretest, but used a different story that was similar in complexity. Each child participated in two 25-minute dynamic assessment sessions. The researchers analyzed four dynamic assessment variables: posttest scores, gain scores, modifiability ratings, and teaching duration. They found that overall modifiability ratings provided the best classification of disorder, with 100% sensitivity and 88% specificity after one teaching session and 100% sensitivity and specificity after two sessions. Any combination of two scores (posttest, modifiability rating, and teaching duration) produced sensitivity and specificity rates over 90%.

Although the research evidence is promising in support of English narrative-based dynamic assessments, more research is needed to cross-validate recent research findings with independent, larger samples of students. The purpose of this study was to examine and cross-validate how well posttest and modifiability scores from the Petersen et al. (2017) dynamic assessment of language (the DYMOND) identify bilingual English/Spanish-speaking students with and without language disorder. The following research questions were explored:

1. To what extent do the dynamic assessment modifiability variables, when added to the dynamic assessment posttest variable, account for variance (R^2) in language ability?
2. What is the optimal sensitivity and specificity of the dynamic assessment?

Method

Participants

This study included 110 diverse school-age (K-6) children, of which 12% (13) had a language disorder (Table 1). Participants were recruited from school districts in Utah and Colorado and were bilingual or monolingual English speakers. Additional demographic information was unavailable. The Petersen et al. (2017) dynamic assessment of language was administered to all the participants.

A-Priori Language Disorder

The children will be identified as having a language disorder if they meet three of the following five criteria: (a) an active IEP for language, (b) parent concern of language development, (c) teacher concern of language development, (d) 70% or less accurate syllables in a nonword repetition task, and (e) 1.5 standard deviations (local norms) or lower on a narrative language task. The children will be identified as not having a language disorder if they do not have an IEP for language and if they do not meet three of the four remaining criteria. Each child was sent home with a permission form on which the child's parents indicated whether or not they had concerns about the child's language development. Each child's teacher was given a survey in which the teacher marked whether or not he/she was worried about the child's language (e.g., reading, writing, speaking) abilities.

Table 1

Demographic Information

			Typically Developing	Language Disorder
		n (%)	*n* (%)	*n* (%)
Total Number of Students		110	97 (88%)	13 (12%)
Gender	Male	52 (47%)	42 (38%)	10 (9%)
	Female	58 (53%)	55 (50%)	3 (3%)
Grade Level	K	3 (2%)	2 (2%)	1 (1%)
	1	21 (19%)	14 (13%)	7 (6%)
	2	20 (18%)	18 (16%)	2 (2%)
	3	21 (19%)	21 (19%)	0 (0%)
	4	13 (12%)	12 (11%)	1 (1%)
	5	21 (19%)	19 (17%)	2 (2%)
	6	11 (11%)	11 (10%)	0 (0%)
Language Diversity	Monolingual English	77 (70%)	70 (64%)	7 (6%)
	Bilingual	33 (30%)	27 (25%)	6 (5%)
Location	Colorado	46 (42%)	41 (37%)	5 (5%)
	Utah	64 (58%)	56 (63%)	8 (7%)

Measures

All participants were assessed in English. Graduate and undergraduate research assistants administered the tests in quiet rooms in the students' schools. Most testing was completed in one day; however, when necessary, testing was completed over the course of two days to mitigate fatigue and to accommodate the students' schedule. For most children, the entire battery of testing required about 30 minutes.

Participants were evaluated using the Narrative Language Measures (NLM), a nonword repetition task, and the Dynamic Measure of Oral Narrative Discourse (DYMOND). The NLM was administered first, followed by the administration of the nonword repetition task and the DYMOND. Each of the assessment measures was audio recorded. The examiners were blinded as to whether the students had a language disorder or not.

CUBED: Narrative Language Measures (NLM). The NLM Listening subtest of the CUBED (Petersen & Spencer, 2016) was used to aid in determining language disorder. The NLM Listening is a language assessment and progress monitoring tool which involves the retelling of a brief narrative. The retell provides a language sample which renders information regarding language complexity and inclusion of story grammar elements. Every student, regardless of grade, was administered one NLM Listening story. The examiner read the model story to the child and then asked the child to retell the story. Each child's retell was audio-recorded and scored in real-time based on inclusion of story grammar elements and language complexity targets.

Non-word repetition task. A nonword repetition task was also used to help determine whether a child had a language disorder. A sample of 10 nonwords from the Children's Test of Nonword Repetition (CNRep; Gathercole, Willis, Baddeley, & Emslie, 1994) was administered with two additional researcher-generated nonwords. Specifically, one 2-syllable low complexity nonword was first presented to each student, followed by two 3-syllable low complexity nonwords then a high complexity 4-syllable nonword, a low-complexity 4-syllable, and a high-complexity 4-syllable nonword. The nonword test increased in difficulty thereafter, with the presentation of a low word-like 5-syllable nonword, two high word-like 4-syllable nonwords, and a 5-syllable low word-like nonword (Gathercole et al., 1994). Finally, a 6-syllable researcher-generated nonword “pristorakitional,” and a high-word like researcher-generated 4-syllable nonword “ruderpation” were presented. The students were instructed to listen to the audio-recording and repeat back each word. The students’ responses were recorded and later scored according to the number of correct syllables the students produced. The number of correct syllables was then totaled (the highest possible score being 51). Scores that fell at or below 70% syllables correct (a score of 36 or lower) on the nonword repetition task were noted. This information was used to help determine the likelihood of a student having language disorder.

DYMOND. All students were given the same dynamic assessment of language, the DYMOND, used in Petersen et al. (2017). The dynamic assessment entailed four steps: a pretest, a teaching phase, a set of modifiability rating scales, and a posttest. The dynamic assessment took approximately 10 minutes, but varied in length based on student 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 real-time 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 the 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 to 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 subordination conjunctions typically occurred only 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 max 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).

Test Administration: Fidelity and Inter-Rater Reliability

Fidelity. A team of undergraduate students in the Communication Disorders program at Brigham Young University were trained using several test protocols. Students who were team leaders received extensive training over several hours. The research assistants of the assessment team were given a training which lasted approximately an hour and were certified and approved by a team leader, suggesting that the individual was ready to administer the assessments on their own. Research assistants administered five practice sessions using the modifiability rating form on fellow research assistants playing the role of children with and without language disorder. Each team member was required to demonstrate competence and ability to adhere to and carry out the testing procedures consistently and independently with 100% accuracy. There was always a team leader and/or an experienced, trained individual onsite to monitor adherence to testing procedures. Fidelity was monitored by team leaders while examiners were administering the dynamic assessment.

Inter-rater reliability. Although this study did not examine intra-rater reliability, inter-rater reliability was examined for 25% of the typically developing children's scores and 50% of the scores from the children with language disorder. The children whose tests were rescored were randomly selected using a random number generator. Selected, trained examiners from the team independently performed the rescoring and were blind to the children's language status (i.e., typically developing, disorder). The independent examiners completed the same training requirements as the primary research assistants and demonstrated the ability to administer and score the dynamic assessment accurately. These individuals listened to the audio files corresponding to the chosen children and scored the pretest, modifiability judgement, modifiability total and posttest scores in real-time. The total scores from the trained individuals were compared to the total scores given by the initial examiner. The percent agreement and the range of agreement were analyzed.

Rather than calculate point-to-point inter-rater reliability, inter-rater reliability was calculated for total subtest (i.e., pretest, modifiability total, modifiability judgement, posttest) scores. The interrater reliability of the pretest total score with a possible maximum score of 35 was calculated to be between 78% and 87.5% when a margin of error (+/- 2, +/- 3), was applied. The interrater reliability of the modifiability total score, which had a maximum score of 24, was between 75% and 88%, provided a margin of error (+/- 2, +/- 3). The interrater reliability of the modifiability judgement score, which is on a scale of 0-4, was calculated to be between 86% and 100% when a range of +/-1 and +/-2 was respectively applied. The interrater reliability of the posttest scores, which similar to the pretest had a total score of 35, was calculated to be between 69% and 91%, with a range of +/- 2 and +/-3 respectively.

Results

Data were analyzed using the Statistical Package for Social Sciences (SPSS version 24.0; International Business Machines Corporation, 2016). Logistic regression and receiver operator characteristic (ROC) analyses were conducted in order to determine to what extent dynamic assessment modifiability and posttest variables accounted for variance in language ability and to determine the optimal combination of sensitivity and specificity for the dynamic assessment.

Logistic regression utilizes independent and continuous predictor variables to predict a binary dependent variable. In this study, language ability was the binary dependent variable (i.e., language disorder/no language disorder) and the continuous predictor variables were the dynamic assessment modifiability and posttest scores.

Hierarchical logistic regression was used to determine to what extent dynamic assessment modifiability variables (total modifiability and modifiability final judgment) accounted for the variance in language ability when combined with the dynamic assessment posttest score (Question 1). In the first hierarchical logistic regression model, the posttest variable was entered into the logistic regression first, in Step 1, followed by the modifiability final judgment score, Step 2. Results of Model 1 indicated that the posttest accounted for 44% of the variance alone (Nagelkerke R^2), and that the combination of the posttest and the modifiability final judgment variables accounted for 56% of the variance in language ability (Nagelkerke R^2 =.56).

In the second logistic regression model, as shown in Table 2, the posttest score was entered first in Step 1, followed by the total modifiability score (Step 2), then by the modifiability final judgement score (Step 3). The variables were entered in this second model in this particular order to investigate whether the total modifiability score positively contributed to the prediction model for language ability. The results of Model 2 indicated that the combination of the posttest and total modifiability variables accounted for 54% of the variance in language ability. The addition of the modifiability final judgment score to those two variables significantly affected the model, with all three variables accounting for 58% of the variance in language ability (Nagelkerke $R^2 = .58$, Wald = 2.27, $p < .01$). In order to determine the optimal sensitivity and specificity of the dynamic assessment (Question 2), receiver operator characteristic (ROC) analyses, which provided area under the curve (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, with both the posttest and modifiability total scores combined, and from the Model 2, Step 3 logistic regression analysis, with the posttest, modifiability total, and modifiability judgment combined, were used as the predictor measures in the ROC analyses, with language ability as the criterion measure. Sensitivity and specificity were held at 70% or higher whenever possible. Results indicated good to excellent classification accuracy for both Model 1 and Model 2, with AUC values ranging from .88 to .93, with sensitivity ranging from 85% to 92% and specificity ranging from 77% to 88%.

Table 2

Logistic Regression and ROC Analyses for Narrative Dynamic Assessment Predictor Variables.

Model	Step	Predictor	*Beta*	*Exp*(B)	R^2	ΔR^2	χ^2	Wald	Sens.	Spec.	AUC
1	1	Posttest	-.30	.74	.44		28.52**	19.06	.86	.77	.88
	2	Mod Judge	-1.26	.28	.56	.12	37.18**	4.95	.92	.84	.92
2	1	Posttest	-.30	.74	.44		28.52**	19.06	.86	.77	.88
	2	Mod Total	-30	.74	.54	.10	35.74**	5.53	.85	.88	.92
	3	Mod Judge	-.24	.79	.58	.04	38.81**	2.27	.92	.83	.93

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* ≤ .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.

Discussion

The purpose of this study was to determine the classification accuracy of the dynamic assessment and to identify to what extent dynamic assessment modifiability and posttest variables accounted for the variance in language ability. Hierarchical logistic regression analyses indicated that modifiability scores and posttest scores when used in combination accounted for the most variance in language ability. Receiver operator characteristic analyses using these dynamic assessment variables revealed good to excellent sensitivity and specificity.

These results cross-validate findings from Petersen et al. (2017). The same dynamic assessment used in Petersen et al. (2017) was administered and scored in the current study by a different group of examiners to a larger, independent sample of students. In most cases, predictive studies report results that are optimized for the specific population under investigation. It is important, however, to demonstrate that procedures applied to one sample of students can also be applied to an independent sample. Because this current study replicated results from previous dynamic assessment research, with moderate to high sensitivity and specificity, there can be greater confidence that the specific dynamic assessment procedures used in this study will yield valid results for a greater population of students. This cross-validation also indicates that different clinicians with varying degrees of experience in testing children can obtain sensitivity and specificity at or above 80%.

Clinical Implications

Need for valid language assessments. Obtaining adequate sensitivity and specificity is particularly important because the majority of traditional static assessments have resulted in poor classification accuracy (Denman et al., 2017; Friberg, 2010; Spaulding et al., 2006). The results of this study indicate that dynamic assessment is a promising alternative to traditional NRTs. This classification accuracy aligns with findings from multiple other studies investigating the dynamic assessment of language. The research evidence is mounting that dynamic assessment has superior classification accuracy over traditional NRTs. Tests designed to diagnose disorders need to have adequate sensitivity and specificity. When assessments lack adequate sensitivity and specificity, the chance of misidentifying children with and without disorders increases. This misidentification can result in over or underrepresentation of children, particularly culturally and linguistically diverse populations, in special education. There has been a historical disproportionate representation of culturally and linguistically diverse students in special education. In fact, 13.5% of students in kindergarten through 12th grade receive special education services, but some population subgroups receive higher or lower rates of services than others (Donovan & Cross, 2002; Morgan, Farkas, Hillemeier, & Maczuga, 2017). It is important that proper intervention be provided to students based on their needs, and the best way to correctly identify needs is with a valid testing measure. The results of this study provide additional evidence that dynamic assessment can be a useful tool in validly identifying children with and without language disorder.

Clinician concerns with dynamic assessment efficiency. Although dynamic assessment has yielded better psychometric qualities with accurate classification, dynamic assessment has not been widely implemented. This may be due to a number of factors, including lengthy administration time and lack of standardization.

Historically, dynamic assessment required a lengthy amount of time to administer and, although accurate, was not time-efficient and plausible to use for many clinicians who have a large workload (American Speech-Langauge-Hearing Association, 2018; Peña et al., 2006). However, the dynamic assessment (DYMOND) examined in this study demonstrated similar sensitivity and specificity as found in lengthier, currently available dynamic assessments, yet is able to be administered in approximately 10 minutes. In fact, the dynamic assessment used in this study can be administered in less time than many of the standardized, norm-referenced assessments that are typically administered. This new, efficient dynamic assessment could be clinically useful, as clinicians may be able to receive valid results in a smaller amount of time.

Clinician concerns with reliability. Due to the subjectivity and clinical judgement required in the administrative process of the dynamic assessment, clinicians have been hesitant to implement dynamic assessment. The historical lack of standard procedure in dynamic assessment has raised clinician concern about obtaining reliable results. In the DYMOND, subjective clinical impressions are important; however, there are administration guidelines and correction procedures for the teaching phase which provide an added measure of standardization. Furthermore, when scoring the modifiability section, the examiner uses a carefully constructed modifiability rating scale which quantifies and objectifies clinical impressions of a child's behavior during the teaching phase. These specific procedures, operationally defined and quantifiable behaviors, and explicit use of a specific modifiability rubric helps to guide the clinician and leads to greater standardization. Previous investigations of the dynamic assessment used in this study have found that point-to-point inter-rater reliability has been good to excellent. The findings of this study add to the current body of research by reporting, for the first time, inter-rater reliability of total scores for pretest, posttest and modifiability with fair to good results when accounting for a small margin of error.

Role of dynamic assessment in informing treatment. In contrast to static, norm-referenced assessments, dynamic assessment can provide both diagnostic information and clinically relevant information for treatment. The teaching phase of the dynamic assessment allows the clinician to identify where breakdowns occur for the child and provides an opportunity to identify how much support the child needs to be successful. Because the dynamic assessment used in this study focused on narrative language, functional goals revolving around narrative discourse can be generated. Oral narrative language is replete with academic language features that all students need to produce and understand in order to be successful in school (Petersen et al., 2017). Therefore, by using a narrative-based dynamic assessment, both diagnostic and clinical information can be obtained.

Study Limitations and Future Research

To our knowledge, this study was the largest dynamic assessment of language investigation conducted to date. Although there were 110 participants, including many who were culturally and linguistically diverse, only a small percentage of the sample had a language disorder. In future research, it is planned that 50% of the recruited sample will have a language disorder. This will reduce base rate confounds and provide greater confidence in generalizability to other children with language disorders.

Future research will also aim to recruit a larger sample size from representative locations across the United States, encompassing not only a larger number of children with language disorder, but also children who represent greater diversity of language and culture. The recruitment of a large number of diverse children from across the United States is essential, as many previous studies lack adequate sample sizes for appropriate psychometric evaluations. Recruiting a larger sample of children, particularly of culturally and linguistically diverse children with language disorders, will be of utmost importance to provide sufficient representation of all students, potentially providing additional evidence of validity of the dynamic assessment of language investigated in this study.

References

American Speech-Language-Hearing Association (2018). *Schools Survey: SLP caseload and workload characteristics report*. American Speech-Language-Hearing Association, Rockville, MD.

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*, 292-305.

Colby, S., & Ortman, J. (2015). *Projections of the size and composition of the U.S. population: 2014 to 2060, Current Population Reports*, U.S. Census Bureau, Washington D.C.

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

Donovan, S., & Cross, C. T. (2002). *Minority students in special and gifted education.* Washington, DC: National Academies Press.

Feuerstein, R., Rand, Y., & Hoffman, M.B. (1979). *The dynamic assessment of retarded performers: The learning potential assessment device theory, instruments, and techniques.* Baltimore, MD: University Park Press.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*, 77-92. doi:10.1177/0265659009349972

Gathercole, S. E., Willis, C. S., Baddeley, A. D., & Emslie, H. (1994). The children's test of nonword repetition: A test of phonological working memory. *Memory*, *2*, 103–127. doi: 10.1080/09658219408258940

Hasson, N., & Joffe, V. (2007). The case for dynamic assessment in speech and language therapy. *Child Language Teaching and Therapy, 23*, 9-25. doi:10.1177/026565900707214

International Business Machines Corporation. Released 2016. IBM SPSS Statistics for Windows, Version 24.0. Armonk, NY: IBM Corp.

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, 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. doi:10.1044/0161-1461(2003/005)

Miller, L., Gillam, R. B., & Peña, E. D. (2001). *Dynamic assessment and intervention: Improving children's narrative skills*. Pro-Ed.: Austin, TX.

Morgan, P.L., Farkas, G., Hillemeier, M/M., & Maczuga, S. 2017. Replicated evidence of racial and ethnic disparities in disability identification in US schools. *Educational Researcher*, 46, 305-322.

Patterson, J. L., Rodríguez, B. L., & Dale, P. S. (2013). Response to dynamic language tasks among typically developing Latino preschool children with bilingual experience. *American Journal of Speech-Language Pathology*, *22*, 103–112. doi: 10.1044/1058-0360(2012/11-0129)

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*, 2208-2220. doi:http://dx.doi.org/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., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *Journal of Special Education,* 26, 269–280.

Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*, 138-154. doi:10.1044/1058-0360(2001/014)

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.

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities*, 48, 3–21. Retrieved from https://doi.org/10.1177/0022219413486930

Petersen, D.B. & Spencer, T.D. (2016). Using narrative intervention to accelerate canonical story grammar and complex language growth in culturally diverse preschoolers. *Topics in Language Disorders, 36*, 6-19. doi: 10.1097/TLD.0000000000000078

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). San Antonio, TX: Pearson Education Inc.

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. doi:10.1044/0161-1461(2006/007)

Ukrainetz, T. A., Cooney, M. H., Dyer, S. K., Kysar, A. J., & Harris, T. J. (2000). An investigation into teaching phonemic awareness through shared reading and writing. *Early Childhood Research Quarterly, 15*, 331-355.

U.S. Census Bureau (2019). *Distribution of race and Hispanic origin by age groups*. Retrieved from https://www.census.gov/library/visualizations/2019/comm/age-race-distribution.html

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

Westby, C. (1985). *Learning to talk—talking to learn: Oral literate language differences.* San Diego, CA: College Hill Press.

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

APPENDIX A

Annotated Bibliography

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*, 292-305.

Objective: The purpose of this study was to evaluate the performance of a group of Spanish-speaking dual language learners who came from low SES and low parent education backgrounds on the CELF-4S.

Method: 656 Spanish-speaking dual language learners were assessed for presence of language disorder using the core language score of the CELF-4S and the English Structured Photographic Expressive Language Test. 299 of the participants were evaluated using a Spanish language sample for identification of language disorder.

Results: Over 50% of the sample scored more than 1 SD below the mean on the core language subtest. In the study sample, the sensitivity of the CELF-4S was 94% and specificity was 65% when using a cutoff score of 85. When using a cut off score of 78, sensitivity was 86% and specificity was 80%, which is the minimum to be deemed adequate by Spaulding, Plante, and Farinella (2006).

Relevance to current work: Even Spanish versions of popular standardized norm-referenced tests have room for improvement. There needs to be improvement in these assessments or an alternative all together.

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

Objective: The objective of this systematic review was to 1) examine the psychometric quality of currently available comprehensive language assessments for school-aged children and 2) to identify which of the reviewed assessments have the best psychometrics quality and evidence for use.

Method: The researchers searched 5 data bases and reviewed relevant websites and textbooks to collect and identify language assessments. The selection criteria they used included 1) standardized, norm-referenced spoken language assessments in English with normative data for English speaking school age students. Only the most recent editions of each assessments were included. Initially found 76 assessments that met such criteria. To further narrow the number of assessments, they added additional exclusionary criteria including 1) assessment being excluded if they were not published within the previous 20 years or 2) if they were screeners or tests of academic achievement and is they were designed to assess language skills across at least 2 of the three domains of spoken language (semantics, syntax/morphology, and discourse). Following the application of the exclusionary criteria, the researchers were left with 15 language assessments on which they conducted their systematic review. The quality of each study or assessment was evaluated using the COSMIN checklist which provided a system for rating the following psychometric properties: internal consistency, reliability, measurement error, content validity, structural validity, hypothesis testing, cross-cultural validity, criterion validity and responsiveness. The ratings were on a scale of excellent, good, fair and poor.

Results: No assessments presented with evidence of structural validity, internal consistency or error measurements. In general, all 15 assessments were found to have limitations in regard to evidence of psychometric quality (validity and reliability).

Relevance to current work: Lack of reporting these data is a serious flaw because it doesn't allow for understand and awareness of possible classification errors. Evidence regarding diagnostic accuracy is lacking. Improvements in methodological quality and reporting of studies is needed to provide evidence for SLPs is understanding diagnostic utility of available assessments. The review identified about 5 assessments which currently presented with better evidence of psychometric quality than others, but substantially more data is needed to show that any of those assessments have "good" psychometric quality. Further research is needed to determine a form of assessment which yields good evidence of psychometric quality. Of the 15 assessments evaluated, the Assessment of Literacy and Language, the Clinical Evaluation of Language Fundamentals-5th edition, the Clinical Evaluation of Language Fundamentals-Preschool: 2nd Edition, and the Preschool Language Scales-5th Edition presented with the most evidence.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*, 77-92. doi:10.1177/0265659009349972

Objective: The objective of this study was to determine the psychometric validity of various language assessments in diagnosing children with and without language disorders.

Method: Only tests which were found to adequate levels of sensitivity and specificity (>80%) were included in this study. Used the criteria procedure from Spaulding et al (2006). The criteria were 1) test purpose was identification of language disorder 2) test was not a screening tool and

3) information related to identification accuracy needed to be provided in the examiner's test manual. 10 language assessments were identified for initial review. Those included: CELF-4, CELF-P2, PEST, PLS-4, SPELT-3. SPELT-P2, TEEM, TEGI, TLC-E and TNL (see page 80 (4) of pdf). The PEST was excluded from the study because it was unable to be located after extensive searching. Thus, 9 assessments were reviewed in this study. 11 criteria were applied to each of the selected tests in order to determine each test's overall level of psychometric validity. The criteria were 1) purpose of the test was identified, 2) test qualifications are explicitly stated, 3) testing procedures are sufficiently explained, 4) adequate standardization sample size (>100) is noted, 5) clearly defined standardization sample including information related to geographic representation, SES, parent education, gender distribution, ethnic background, presence/absence of impairment and age, 6) evidence of item analysis, 7) measures of central tendency are reported, 8_ concurrent validity is documented, 9) predictive validity is documented, 10) test/re-test reliability is reported, and 11) inter-rater reliability is reported.

Results: Nine preschool and school-age language assessments were found to have acceptable levels of classification accuracy. Of the 9 assessments evaluated, no assessment tool met all 11 criteria. Across the 9 assessments, tests ranged from fitting 8 to 10 criteria. Each of the selected assessments met at least 8 of the 11 criteria. Five tests met 10 out of the 11 criteria. Most assessments had adequate levels of each criteria. However, 78% did not meet the predictive validity criteria and 44% did not meet the test-retest criteria.

Relevance to current work: While there are assessments with adequate psychometrics, there has to be something better. Norm referenced tests should also only be used with children whose demographics are represented in the normative sample, which may not be the case in many of these assessments.

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*, 119-128.

Objective: The purpose of this study was to examine the classification accuracy of the Dynamic Assessment and Intervention (DAI) tool created by Miller, Gillam and Peña in 2001.

Method: The DAI was given to 17 third-grade children from a First Nations community. Each child was determined to have normal language learning skills or possible language learning difficulty. Each child was administered the DAI.

Results: Both the normal language learning skills and the possible language learning difficulty groups benefitted from the teaching phase where they were taught specific targets. The children in the normal language learning skills group, however, demonstrated greater benefit and gains and also generalized more often to targets that were not specifically addressed in the teaching phase. Discriminant analyses revealed high sensitivity and specificity.

Relevance to current work: The DAI and dynamic assessment in general can be a useful tool is accurately diagnosing language disorders in the First Nations community and potentially other populations.

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. doi:10.1044/0161-1461(2003/005)

Objective: The purpose of this forum was to describe and propose dynamic assessment as an alternative assessment method for identifying language disorders in culturally and linguistically diverse population.

Method: This forum outlined the increasing diversity in the country and the limitation of standardized norm-referenced tests when used on these populations. Content and linguistic biases are discussed as well as disproportionate representation in normative samples.

Relevance to current work: Dynamic assessment is an appealing and promising alternative to NRTs because they focus on an individual's ability to learn and not their current knowledge and experience. Furthermore, they can be quick and easy to administer. Further research needs to be done with dynamic assessment as an alternative testing procedure for identifying language disorder in CLD children.

Lidz, C. S., & Peña, E. D. (1996). Dynamic asessment. *Language, Speech, and Hearing Services in Schools, 27*, 367-372. doi:10.1044/0161-1461.2704.367

Objective: The purpose of this article was to propose dynamic assessment as an alternative means of language assessment.

Method: The researchers outlined an adaptation of dynamic assessment as a diagnostic measure for identifying language disorder in Latino preschool children.

Results: Dynamic assessment yields promising results regarding the classification accuracy of preschool Latino children with and without language disorders.

Relevance to current work: Dynamic assessment is a promising alternative to current assessment methods.

Patterson, J. L., Rodríguez, B. L., & Dale, P. S. (2013). Response to dynamic language tasks among typically developing Latino preschool children with bilingual experience. *American Journal of Speech-Language Pathology, 22*, 103-112. doi:10.1044/1058-0360(2012/11-0129

Objective: The purpose of this study was to whether typically developing, bilingual preschool students demonstrated learning when given a dynamic assessment in a test-teach-retest format.

Method: Three dynamic assessments were given to 32 typically developing 4-year-olds using a graduated prompting approach in the teaching phase. 16 children were Spanish-dominant and 16 children with English dominant. The dynamic assessment was administered in the language to which the child had the most exposure. The three dynamic assessments were based on word learning, semantics and phonological awareness.

Results: The researchers found that children performed significantly higher the final test items as compared to the initial test items for the semantic and word learning tasks, assuming the child required a cue on item 1. There was not a significant difference, however, between the initial and final items on the phonological awareness task. These findings suggest, at least for some tasks, that increase in performance results from graduated prompting, which may be an indicator of modifiability.

Relevance to current work: Graduated prompting, as used in the teaching phase of dynamic assessment, facilitates improve performance and suggests modifiability.

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*, 2208-2220. doi:http://dx.doi.org/10.1044/2014_JSLHR-L-13-0151

Objective: The purpose of this study was to evaluate the classification accuracy of a narrative dynamic assessment of language for school-age students learning English as a second language.

Method: The narrative dynamic assessment was administered to 54 children, 18 of which were Spanish-English-speaking with language impairment, 18 of which were typically developing controls matched by age, sex IQ and language experience, and lastly 18 were in a comparison group, matched for age and language. Scores were taken for pretest, modifiability, and posttest.

Results: Discriminant analyses revealed that a combination of modifiability ratings, story scores and ungrammaticality from a language sample yielded 80.6% to 97.2% classification accuracy.

Relevance to current work: Dynamic assessment yields good classification accuracy for students learning English as a second language. Furthermore, modifiability scores are predictive of language disorder.

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 to examine the reliability and classification accuracy of a narrative-based dynamic assessment of language. The first research questions dealt with the extent to which the two stimulus stories were parallel. The second question dealt with the difference in performance between children with language impairment and typically developing counterparts.

Method: 58 first and second grade children were asked to narrate two stories using wordless picture books. The stories were then rated on various aspects of language form and content. Then, 71 children were given the dynamic assessment in a test-test-retest format. The children were asked to narrative a story provided a wordless picture book at pretest. Then, they went through a teaching phase where they received explicit instruction regarding narration. The children were then asked to narrate a new story at posttest.

Results: The researchers found that the measures applied to their pre and posttest stories reflected good internal consistency. During the dynamic assessment portion, the researchers found that typically developing students who participated in the teaching phase demonstrated greater gains from pre to posttest than the students in the language impairment and control groups. Sensitivity and specificity values were highest where modifiability and posttest scores were used as predictors of language ability. Specifically, the combination of modifiability and posttest scores resulted in perfect classification accuracy.

Relevance to current work: Dynamic assessment is a promising method of assessment for accurately identifying language impairment in school-age children. Modifiability and posttest scores demonstrated the highest classification accuracy when used a predictors of language ability.

Peña, E., Iglesias, A., & Lidz, C. S. (2001). Reducing Test Bias Through Dynamic Assessment of Children's Word Learning Ability.

Objective: The purpose of this study was to examine the performance of CLD preschool children on a word-learning task and to determine the classification accuracy of a dynamic assessment as compared to a static assessment.

Method: This study consisted of 79 preschool-age children who were enrolled in a bilingual Head Start program. Each child was administered a dynamic assessment which followed a test-teach-retest model. The children were divided into two groups, with one group receiving mediation and the other group no mediation. Those children in the mediation group were taught naming strategies during a teaching phase.

Results: Children who had typical language development were differentiated from those who had low language skills when using the dynamic assessment process. The dynamic measures such as posttest scores and modifiability ratings had high classification accuracy (sensitivity and specificity) than static measures such as pretest scores. The findings suggest that dynamic assessment is effective is differentiating difference vs disorder.

Relevance to current work: Dynamic assessment yields high classification accuracy, especially for CLD students. Posttest and modifiability scores are most predictive of language ability.

Peña, E., Quinn, R., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *The Journal of Special Education, 26*, 269-280. http://dx.doi.org/10.1177/002246699202600304

Objective: The purpose of this study was to demonstrate the application of the mediated learning experience to language assessment and to determine if the classification accuracy of this dynamic method was more favorable than static measures.

Method: This study has 50 participants who were mostly of Puerto Rican and African American descent and most spoke Spanish and English. Received classroom instruction in both languages. Pretest-teach-posttest method was employed. At pretest, the children were administered the EOWPVT. There was then a mediation (teaching) phase which consisted of two 20 minute sessions in small groups. The examiners then recorded the results of the mediated learning experience by using a modified version of the Dynamic Assessment Recording Form (Lidz, 1991). The recording form helped to guide and record information regarding the children's ability to attend, self-regulate and use the adult/examiner as a resource. Then modifiability was rated using three Likert-type scales from Lidz's (1991) Summary of Dynamic Assessment Results. The rating scales addressed the child's responsiveness, examiner effort and transfer between teaching phases. Following the teaching phase, the EOWPVT was administered again at the end of the year as a posttest.

Results: The use of a dynamic method better differentiates between children with and without language disorders (also difference vs disorder) than static measures. PLD (possibly language disordered) children, on overage, were less responsive to mediation and required more effort on the part of the examiner to induce change. The results validate the underlying construct on dynamic assessment and confirmed their hypothesis that PLD children were less modifiable.

Relevance to current work: Dynamic forms of assessment are more valid and accurate at identifying langue disorders, particularly for children who are culturally and linguistically diverse. Furthermore, static forms of assessment are consistently poor at identifying disorders and differentiating between difference and disorder in CLD children.

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.

Objective: The purpose of this study was to identify the classification accuracy of an English dynamic assessment of narrative language in identifying language disorders

Method: This study consisted of 42 Spanish-English bilingual K-3 children, 1/3 of which had language disorders. Each child was administered two 25-miuntes dynamic assessments, which took a pretest-teach-posttest model. The pretest and posttest consisted of the retelling of narratives, which were scored in real time. Between the pre and posttests, there was a teaching phase in which structured intervention was provided. Examiners taught children missing story grammar elements and language complexity targets. The pretest and posttests were identical in administration, but each had a different target story which was parallel in structure and difficulty.

Results: Of the 4 dynamic assessment variables that were analyzed (posttest scores, gain scores, modifiability ratings, and teaching duration), discriminant function analysis revealed that overall modifiability ratings were most predictive of language disorder. When using the modifiability ratings as predictors of language disorder, there was 100% sensitivity and 88% specificity after the first DA session and 100% sensitivity and specificity after the second DA session. Any two combination of posttest scores, modifiability ratings and teach duration after a single session resulted in 90% sensitivity and specificity across the board. A post hoc question lead to the finding that similar classification accuracy can be yielding with a single 5-10 minute teaching cycle, rather than two, which further abbreviates the DA process.

Relevance to current work: Dynamic assessment is an accurate form of language assessment for CLD students. Furthermore, modifiability scores or a combination of both yield high sensitivity and specificity. Gain scores may not be not be very predictive or helpful. The DA can be further abbreviated and maintain high classification accuracy.

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities*, 48, 3–21. https://doi.org/10.1177/0022219413486930

Objective: The objective of this study was to determine the predictive validity of a dynamic assessment which was designed to identify later risk for reading difficulty and language disorder in bilingual Spanish-English-speaking children.

Method: 63 kindergarten bilingual Spanish-English speaking children completed a dynamic assessment of nonsense words. The DA took a test-teach-retest format. At the end of first grade, the same participants completed a criterion measure of word identification and decoing.

Results: The dynamic assessment yielded high classification accuracy, with sensitivity and specificity at or above 80% for each reading measure, including 100% sensitivity for 2/3 first-grade measures.

Relevance to current work: The model of dynamic assessment has high sensitivity and specificity and has excellent accuracy in determining future language difficulty/performance.

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.

Objective: The purpose of this study was to evaluate psychometric quality of the Preschool Language Scale-3: Spanish Version and its utility in the language assessment of Spanish/English-speaking children.

Method: The authors applied 10 psychometric criteria establish by McCauley and Swisher (1984) and 5 additional criteria from Hutchinson (1996). McCauley and Swisher's criteria were 1) description of normative sample, 2) sample size, 3) item analysis, 4) means and standard deviations, 5) concurrent validity, 6) predictive validity, 7) test-retest reliability, 8) inter-examiner reliability, 9) description of test procedures, and 10) description of examiner qualifications. The additional 5 criteria from Hutchinson were 1) purpose of the test explicitly stated, 2) construct or model explicitly stated, 3) supportable rationale for test content, 4) sample behavior at the extremes and 5) norms represent performance at extremes. 37 bilingual Spanish-English speaking children (18 boys and 19 girls). They were administered the PLS-3S.

Results: The evaluation revealed that the PLS-3S met only 4/10 criteria proposed by McCauley and Swisher and none of the additional criteria from Hutchinson. There were problems identified in the test's norming and in the lack of reliability and validity data. Children's performance, even as TD children, was approx. 1.5 SD below the man. And the performance on subtests did not reflect an even progression of item difficulty, indicating limited evidence of construct and content validity.

Relevance to current work: Even Spanish versions of popular standardized norm-referenced tests have problems. There needs to be improvement in these assessments or an alternative.

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. doi:10.1044/0161-1461(2006/007)

Objective: The objective of this study was to determine the classification accuracy of 43 commercially available norm-referenced language assessments.

Method: Data from the test manuals of 43 commercially available norm-referenced language assessments were compiled to identify psychometric quality and classification accuracy for identifying language disorders in children.

Results: A review of the test manual data revealed that children with language disorders did not consistently score at the low end of the assessment's normative distribution. A majority of the reviewed assessments reported that some of the children with language disorders score above 1.5 SD below the mean and that in 27% of the assessments, there were children with language disorders who had scores which fell within 1 SD of the mean. Furthermore, only 9 of the 43 reviewed assessments reported sensitivity and specificity. Of those 9 assessments which reported sensitivity and specificity, only 5 of the 9 assessments had acceptable accuracy (>80%).

Relevance to current work: Norm-referenced tests are most commonly used to identify language disorders in children. The majority of these assessments, however, lack high classification accuracy and are not consistent in correctly identifying language disorders in children. The lack of classification accuracy results in the misidentification of children with and without language disorders. These findings suggest that there is a need to develop language assessment measures with adequate levels of sensitivity and specificity.

Ukrainetz, T. A., Cooney, M. H., Dyer, S. K., Kysar, A. J., & Harris, T. J. (2000). An investigation into teaching phonemic awareness through shared reading and writing. *Early Childhood Research Quarterly, 15*, 331-355.

Objective: The purpose of this study was to examine a dynamic assessment of word learning skills and to identify whether it yields good classification accuracy in determining language disorder in bilingual children.

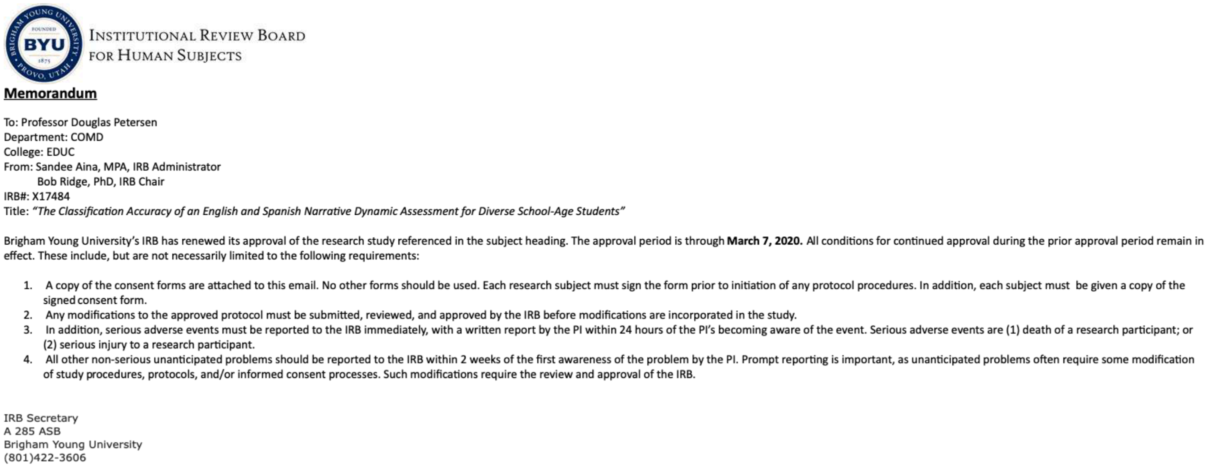
Method: 15 predominantly Spanish-speaking children between the ages of 4 and 5 with typically developing language skills and 13 with primary language impairment participated in a 30-40 minute dynamic assessment, which took the form of a test-teach-retest design.

Results: The researchers found that the typically developing group made associations between phonological and semantic representations of new words faster than those children in the primary language impairment group. Typically developing children demonstrated better modifiability than the language impairment group.

Relevance to current work: A dynamic assessment of word learning may accurately identify language impairment and differentiate typically developing language ability from disorder.

APPENDIX B

IRB Approval



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