2021-04-27

Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Large-Scale Psychometric Study

Anahi Kamila DeRobles
Brigham Young University

Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Large-Scale Psychometric Study

Anahi Kamila DeRobles

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
Connie Summers
Tyson G. Harmon

Department of Communication Disorders

Brigham Young University

ABSTRACT

Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Large-Scale Psychometric Study

Anahi Kamila DeRobles
Department of Communication Disorders, BYU
Master of Science

The purpose of this study was to examine and cross-validate how well a dynamic assessment of language can accurately identify a large sample of school-age students with a representative ratio of language disorder. The participants included 362 school-age children with and without language disorder from kindergarten to sixth grade in Utah, Colorado, and Wyoming. Each participant received a battery of assessments including a dynamic assessment of narrative language. 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. The results of this study indicate that the DYMOND may be a valid and accurate tool when identifying language disorders in school-age populations.

Keywords: language, narratives, dynamic assessment, school-age, diverse students

ACKNOWLEDGMENTS

The author would like to thank the following committee members for their meaningful contribution: Douglas B. Petersen, Connie Summers, and Tyson G. Harmon. 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

LIST OF TABLES

LIST OF FIGURES

DESCRIPTION OF THESIS STRUCTURE

To adhere to traditional thesis requirements and journal publication formats, this thesis, *Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Large-Scale Psychometric Study*, 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 first-grade spring benchmark story 1 that was used to help identify language disorder. The Institutional Review Board (IRB) approval is included in Appendix C.

Introduction

There is an increasing need to develop valid and reliable language assessments, particularly for school-age children who are culturally and linguistically diverse. By 2044 more than half of the U.S. population is projected to belong to a minority group (Colby & Ortman, 2015). Additionally, the Spanish speaking population in the U.S. is rapidly growing. In 1980, those who were ethnically Hispanic made-up 9 percent of the population, while now it is estimated to be at 26 percent (Vespa et al., 2018). The 2018 U.S. Census Bureau report estimated that non-Hispanic white residents only make up 49.9 percent of the population under 15 years old (Vespa et al., 2018). According to the U.S. Census Bureau, the Hispanic population will continue to increase, more than doubling by the year 2060 (Vespa et al., 2018). With this projected growth in diversity, school-age language assessments will need to have strong evidence of validity for a diverse population.

Norm-Referenced Language Tests

Speech-language pathologists (SLPs) often rely solely on norm-referenced tests (NRTs) when assessing and diagnosing children with language disorders and do not routinely use other means like language sampling (Caesar & Kohler, 2009). For example, Pavelko et al. (2016) surveyed 1,399 school-based SLPs and found that only 67% of clinicians use informal measures, such as a language sample, to diagnose disorder. Similarly, Williams and McLeod (2012) in their survey found that 80% of SLPs did not use informal measures for diagnostic purposes, yet most reported the use of NRTs. Additionally, Betz et al. (2013) reported that nearly 100% of the SLPs they surveyed ranked NRT's as one of the top 5 most important measures and 50% of SLPs rated NRTs as the most important diagnostic measure.

While these NRTs play a prominent role in language assessments, they often lack adequate sensitivity and specificity (e.g., 80% or higher) for identifying language disorders in school-age children (Spaulding et al., 2006). *Sensitivity* refers to a test's ability to correctly identify children with a language disorder while *specificity* refers to a test's ability to correctly identify children without a language disorder. A study by Spaulding et al. (2006) reviewed the classification accuracy using sensitivity and specificity ratings and the mean group difference in 43 commonly used NRT's for language. Of the 43 tests only 9 reported sensitivity and specificity information and only 5 of those had adequate classification accuracy. Their study also emphasized that an evidence-based practice framework for diagnostic accuracy is measured primarily through sensitivity and specificity data. Currently, most commonly used NRT's lack adequate sensitivity and specificity or do not report those metrics. In a similar study conducted by Friberg (2010), nine preschool and school-age NRT's were evaluated based on an identification accuracy of .80 or better with specific psychometric criteria. The tests included CELF-4, CELF-P2, PLS-4, SPELT-3. SPELT-P2, TEEM, TEGI, TLC-E and TNL. While all 9 tests reported acceptable psychometric criteria, no assessment met all 11 parameters and 7 out of 9 assessments did not meet the predictive validity criteria or did not report it. These data further indicate current deficits in commonly used NRT's and the lack of valid and reliable measures for accurate identification.

Although the classification accuracy of many NRTs fail to successfully identify school-age children, NRTs have statistically lower classification accuracy for culturally and linguistically diverse (CLD) students. This lack of accurate classification for CLD students can be explained by NRTs use of static measures to identify students, which only assess a student's current knowledge and not their capability for learning. NRTs can also include biased test items,

materials, and procedures. Laing and Kamhi (2003) outlined three main factors that contribute to bias: content bias, linguistic bias, and the disproportionate representation of ethnicities in normative samples. Content bias refers to test items that assume cultural and life experiences (e.g., exposure to vocabulary, early literacy experience, and teacher to student traditions). Linguistic bias refers to a language or dialect difference with the student and examiner/test. A language or dialect difference may incorrectly identify a student with a language disorder when the student may have typical language in their first language or dialect. Additionally, NRT's often exclude or under-represent CLD students from their normative samples. This underrepresentation results in a normative sample that is not reflective of CLD students.

Dynamic Assessments of Language

While NRT's tend to lack sensitivity and specificity, research indicates that dynamic assessment has strong classification accuracy, especially for CLD students. Therefore, dynamic assessment is a promising alternative to NRTs. In contrast to NRTs static and fixed measures, dynamic assessments measure a student's ability to learn. One common approach to dynamic assessment includes a pretest-teach-retest model (Peña & Iglesias, 1992). During the pretest phase an examiner measures a student's current ability to perform a certain task. The teaching phase entails systematic and explicit instruction targeting language features. The posttest measures the student's ability to independently integrate the language features taught in the teaching phase. A student's learning potential, or modifiability is revealed during the teaching phase (Feuerstein et al., 1979). Modifiability is often measured using a modifiability rating scale that accounts for how difficult it was for a student to learn and how much effort the examiner had to expend to teach the child (Peña et al., 2006). This unique teaching phase and focus on

modifiability mitigates test confounds associated with cultural and linguistic diversity, including limited English language proficiency, socioeconomic status, or other cultural differences.

Several studies have investigated the role of dynamic assessments of language in classifying language disorder. For example, Peña and Iglesias (1992), Ukrainetz et al. (2000), and Kapantzoglou et al. (2012) investigated the classification accuracy of various dynamic assessments of vocabulary. Additionally, Petersen and Gillam (2015) studied the role of dynamic assessment of reading in identifying children with reading disorders. However, narrative language has been one of the primary focuses of dynamic assessment of language research, as narratives can be highly effective measures of language ability. The use of narratives allows for engaging, rich contexts that measure integrated, academic language in a naturalistic setting (Ukrainetz et al., 2000; Westby, 1985). These benefits are not commonly attainable in NRTs due to their systematic and decontextualized approach. Narratives, on the other hand, contain story grammar elements including character, setting, problem, emotion, action, consequence and ending. These components integrate narrative and academic language use including tier 1, 2, and 3 vocabulary (Beck et al., 2002), language complexity, and subordinate clauses (e.g., adverbial, nominal). Overall, narrative language requires a clear and efficient use of complex language elements.

A growing body of evidence demonstrate the validity of using narrative-based dynamic assessments for accurately identifying language disorders in diverse populations (e.g., Hasson & Joffe, 2007; Peña et al., 2001; Patterson et al., 2013; Peña & Iglesias, 1992). For example, Kramer et al. (2009), Miller et al. (2001), and Peña et al. (2006) examined the classification accuracy of narrative dynamic assessments and investigated the most predictive test variables for language disorder. Each study followed a pretest, teaching phase, posttest model of assessment.

All of these studies found that dynamic assessments yield high levels of classification accuracy for diverse children with language disorder. These studies concluded that the strongest and best predictors of language disorder are posttest scores and clinician judgement using modifiability rating scales. Although promising, Miller et al. (2001) and Kramer et al. (2009) do not include a large sample of participants, a representative sample, nor have their findings been replicated or cross-validated. Peña et al. (2006) presented a larger sample size of 58 bilingual participants but still requires further replication of findings and a more time efficient administration of tests.

Most recently, Orellana et al. (2019) conducted a systematic review of dynamic assessment of language research conducted with bilingual children. Using a meta-analysis procedure, the authors ultimately examined six additional studies: Roseberry & Connell (1991), Peña (2000), Kramer et al. (2009), Kapantzoglou et al. (2012), Peña et al. (2014), and Petersen et al. (2017). Each article was reviewed for their use of methodological quality indicators for dynamic assessments that included the use of a one-gate design (i.e., blind researchers), participants receiving the same test, independent testing, more than 30 participants, blinded testing, valid and reliable reference standards, fidelity to procedure, highly replicable procedures, and a representative ratio of typically developing students to students with language disorder. Across the various dynamic assessment studies, two key patterns were established as being effective for diagnostic accuracy in CLD populations (a) children with language disorder (LD) performed significantly lower than their typically developing (TD) peers at both pretest and posttest for each language measure and (b) clinician judgments of modifiability scored participants with LD significantly lower scores than their TD control groups. While this review concluded that the use of dynamic assessment of language can lead to more accurate identification of LD in CLD populations, these studies presented with weaknesses.

Only three out of six studies included an adequate sample size (30 or more) with Peña (2000) including 55 participants, Peña et al. (2014) with 54 participants, and Petersen et al. (2017) including 42 participants. Additionally, only two studies included a one gate design, five studies had high replicability of procedures but no studies included a representative ratio.

Prospective studies should address weaknesses identified in previous dynamic assessment studies and in the Orellana et al. (2019) meta-analysis. Specifically, future studies should identify which dynamic assessment variables are most predictive of language ability, include a larger sample size, include representative ratios of children with and without language disorder, use a one-gate design where researchers are blind to language ability prior to testing, and cross-validate and replicate previous findings. The purpose of this study was to addresses weakness in previous dynamic assessment studies by examining and cross-validating how well posttest and modifiability scores from the Petersen et al. (2017) dynamic assessment of language can identify a large sample of school-age students with a representative ratio of 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 in a large sample of school-age students?
2. What is the optimal sensitivity and specificity of the dynamic assessment in a large sample of school-age students?

Method

Participants

Participants in this study include a group of 362 diverse students who primarily represented two major races/ethnicities (white and Hispanic). This group included students with

and without language disorder from kindergarten through sixth grade from four elementary schools in Utah, Colorado, and Wyoming. Language disorder was established at the outset for each participant when all three of the following index measures and corresponding criteria were met: (a) an active individualized education program (IEP) for language services, (b) 70% or less accurate syllables in a non-word repetition (NWR; Dollaghan & Campbell, 1998) task, and (c) a score that is -1.5 standard deviations (using sample-specific data) or lower on a narrative language task (NLM). These criteria included participants who had a language disorder in addition to or secondary to another disability. Information on whether or not a student has a current IEP for language services was obtained from the speech language pathologist at each school. Demographic information about the participants including gender, ethnicity, home language, location, grade, and language status are displayed in Tables 1 and 2.

Table 1

Demographic Information for Total Sample

		n (%)
Number of Students		362
Gender		
	Male	180 (49.7%)
	Female	182 (50.3%)
Ethnicity		
	Caucasian	215 (59.4%)
	Hispanic	123 (34%)
	African American	5 (1.4%)
	Native American	11 (3%)
	Asian American	7 (1.9%)
	Pacific Islander	1 (0.3%)
Grade Level		
	K	7 (1.9%)
	1	41 (11.3%)
	2	77 (21.3%)
	3	52 (14.4%)
	4	77 (21.3%)
	5	67 (18.5%)
	6	41 (11.3%)
Language Status		
	Typically Developing	308 (85.1%)
	Language Disorder	54 (14.9%)
Location		
	Rural	141 (39%)
	Urban	221 (61%)
Linguistic home environment		
	Languages other than English	31 (8.6%)
	English only	331 (91.4%)

Table 2

Demographic Information for Students With and Without Language Disorder

		Typically Developing n (%)	Language Disorder *n* (%)
Number of Students		308	54
Gender			
	Male	146 (47.4%)	34 (63%)
	Female	162 (52.6%)	20 (37%)
Ethnicity			
	Caucasian	177 (57.5%)	38 (70.4%)
	Hispanic	110 (35.7%)	13 (24.1%)
	African American	4 (1.3%)	1 (1.9%)
	Native American	9 (2.9%)	2 (3.7%)
	Asian American	7 (2.3%)	0
	Pacific Islander	1 (0.3%)	0
Grade Level			
	K	2 (0.6%)	5 (9.3%)
	1	25 (8.1%)	16 (29.6%)
	2	63 (20.5%)	14 (25.9%)
	3	47 (15.3%)	5 (9.3%)
	4	68 (22.1%)	9 (16.7%)
	5	62 (20.1%)	5 (9.3%)
	6	41 (13.3%)	0
Location			
	Rural	110 (35.7%)	31 (57.4%)
	Urban	198 (64.3%)	23 (42.6%)
Linguistically home environment			
	Languages other than English	22 (7.1%)	9 (16.7%)
	English only	286 (92.9%)	45 (83.3%)

Measures

The entire battery of assessments includes the Narrative Language Measures (NLM), a non-word repetition task (NWR), and the dynamic assessment of language (the DYMOND). A total of 16 undergraduate and 2 graduate research assistants administered the NWR task, NLM, and the dynamic assessment in English in the Fall of 2018 and the Winter of 2019. Testing was completed in the students' school in approximately 30 minutes when administering the full battery of assessments. To accommodate individual circumstances, school schedules, and possible fatigue, some testing was completed over the course of two days when necessary. Each examiner audio recorded the assessments and were blinded of the student's prior language abilities.

CUBED: Narrative Language Measures (NLM)

The *NLM Listening* subtest of the *CUBED* (Petersen & Spencer, 2012) is a language assessment and progress monitoring tool that was used as an index measure to determine language disorder and provide a language sample for each student. The examiner read the provided script and narrative and then prompted the student to retell that same narrative. The examiner only provided neutral encouragement and scored the students' response in real time. This retell provided information on a student's oral language complexity and ability to include story grammar elements. Each response was audio-recorded for further analysis. The first-grade spring benchmark story 1 was administered to each student regardless of grade. There are a total of 34 points possible. Student scores will be referenced to sample-specific data, and those students who score -1.5 standard deviations or lower from the mean will be identified.

Non-Word Repetition Task (NWR)

The NWR task included a sample of 10 non-words from the Children's Test of Non-Word Repetition (CNRep; Gathercole et al., 1994) and two additional research-based non-words, each word ranging from 2-6 syllables in complexity (Romero, 2019). This task will also serve as an index measure to help determine language disorder. The examiner instructed each student to listen to an audio-recording of the non-words and repeat back each word. Their responses were audio-recorded and later scored based on number of correct syllables out of 51 possible. Students with 36 syllables correct or lower (71% or below) will be identified for language disorder classification.

DYMOND: Dynamic Assessment of Language

The DYMOND is a dynamic assessment of language that includes four distinct steps: a pretest, a teaching phase, a set of modifiability rating scales, and a posttest (Petersen et al., 2017). Each student was administered an English DYMOND that took approximately 10 minutes, depending on the student's responsiveness. The examiner scored the test in real time and audio-recorded the students' responses for further analysis.

Dynamic Assessment Pretest. The pretest of the dynamic assessment includes a scripted narrative read by the examiner and a student retell of that same narrative. The examiner scored the pretest in real time, assessing the student's ability to include each story grammar element (character, setting, problem, feeling, plan, attempt, consequence, feeling-2, plan-2, action-2, consequence-2, ending, and end feeling), as well as elements of language complexity including the use of subordinating conjunctions: because, when, and/or after. The pretest has a maximum score of 35 points, made up of the story grammar subtotal (2 points per element, total of 26 points), and the language complexity subtotal (up to 9 points total).

Dynamic Assessment Teaching Phase. The teaching phase included two distinct steps to facilitate individual learning and promote independent production of narratives with all modeled story grammar elements and language complexity targets. The first step included pictures and icons visually supporting the narrative told at pretest. The examiner presented the set of pictures and icons and pointed to corresponding images while retelling the story. The examiner also provided explicit instruction for each story grammar element and its accompanying picture and icon (e.g., "This is Sam. Sam is the main character of our story. Who is the character?"). The student was then asked to follow the examiners model of retelling the story with the pictures and icons. The examiner provided assistance in including all story grammar elements and including any language complexity targets. The second step included removing the pictures and only using the icons for support. The student then retold the narrative again with only the icons and examiner support when needed.

Examiners followed explicit teaching procedures to provide support for each student to include all the story grammar elements and language complexity targets. When a student omitted story grammar elements or incorrectly retold parts of the narrative, the examiner immediately stopped them and used a level 1 or 2 prompt. Level 1 prompts are open-ended questions (e.g., "Who is this story about?") and were followed by level 2 prompts if not answered correctly. Level 2 prompts require modeling a correct response (e.g., "Sam is the main character of this story.") and then asking the student to repeat it (e.g., "Now you say that"). After either prompt, the examiner used an overcorrection procedure which instructs the student to go back one-story grammar element and resume the retell with the previously omitted element (e.g., "Great! You told me the feeling! Start telling the story again at the problem. Remember to tell me about the feelings"). The examiner had the option to promote the use of language complexity through

increased use of subordinating conjunctions if the student was able to produce all or most of the story grammar elements with ease.

Dynamic Assessment Modifiability. The DYMOND modifiability ratings include a modifiability score and a final modifiability examiner judgment score using a 5-point Likert scale. Immediately after the teaching phases the examiner rated the student's modifiability, or learning potential, and reflected on how difficult it was for the student to learn and how much effort it took to teach the student (Peña et al., 2006). The modifiability rating consisted of the following six criteria using the 5-point scale (24 points possible): response to prompts, degree of transfer, attention to teaching, ease of teaching, frustration, and disruptions. The final modifiability examiner judgment score reflects the overall ease in learning with 4 representing considerable ease and 0 representing significant difficulty.

Dynamic Assessment Posttest. The DYMOND posttest includes a different story than the pretest and teaching phase but was controlled in structure and complexity (e.g., story length, use of tier-two words, dual-episode story structure, inclusion of subordinate clauses). The administration and scoring for the posttest were the same as the pretest.

Test Administration: Fidelity and Inter-Rater Reliability

Fidelity

A team of graduate and undergraduate students in the Communication Disorders program at Brigham Young University were trained using several test protocols. Four students, who were team leaders, received extensive training over several hours and trained the remaining research assistants on the team. Subsequent training lasted approximately one hour and required administration of three practice sessions using the entire battery of assessments to a team leader. Each team member was required to demonstrate competence of the testing procedures and 100%

accuracy in the administration of the battery of assessments. Team leaders authorized training completion before allowing independent administration. Fidelity was monitored by team leaders while examiners were administering the dynamic assessment in real time.

Inter-Rater Reliability

Inter-rater reliability was calculated for approximately 7% of the typically developing children and calculated for 7% of the children with language disorder. The children whose tests were rescored were randomly selected using a random number generator. Inter-rater reliability was examined by selected, trained examiners from the team and were blind to whether or not a student has a language disorder. Students were randomly selected using a random number generator. The independent examiners listened to the audio files corresponding to the selected children and score 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 will be analyzed.

The interrater reliability of the pretest total score with a possible maximum score of 35 was calculated to be 31%. However, when a range of scores was applied (+/- 2, +/- 3), reliability increased to 78% and 87.5 % respectively. The interrater reliability of the modifiability total score was calculated to be 28%. When a range of scores was applied (+/- 2, +/- 3), interrater reliability increased to 75% and 88% respectively. The maximum score for total modifiability was 24. The interrater reliability of the modifiability judgement score, which is on a scale of 0-4, was calculated to be 60%, increasing to 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 25%, increasing to 69% and 91% with a range of +/-2 and +/-3 respectively.

Results

Data were analyzed using the Statistical Package for Social Sciences (SPSS version 27.0; IBM Corporation, 2020). Binary 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 DYMOND. A Binary 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 (step 1), followed by the modifiability final judgment score (step 2). As shown in Table 3, results of model 1 indicated that the posttest accounted for 40% of the variance alone (Nagelkerke R^2), and that the combination of the posttest and the modifiability final judgment variables accounted for 70% of the variance in language ability (Nagelkerke R^2 =.56).

In the second logistic regression model, the posttest score was entered first (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 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 66% of the variance in language ability. The addition of the modifiability final judgment score to those two variables did not increase the R^2, with both model 1 and 2 accounting for 70% of the variance in language ability (Nagelkerke $R^2 = .70$).

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 80% or higher. As shown in Figure 1, results indicated good to excellent classification accuracy for both model 1 and model 2. Results show area under the curve values ranging from .87 to .95, with sensitivity ranging from 83% to 93% and specificity ranging from 76% to 83%.

Table 3

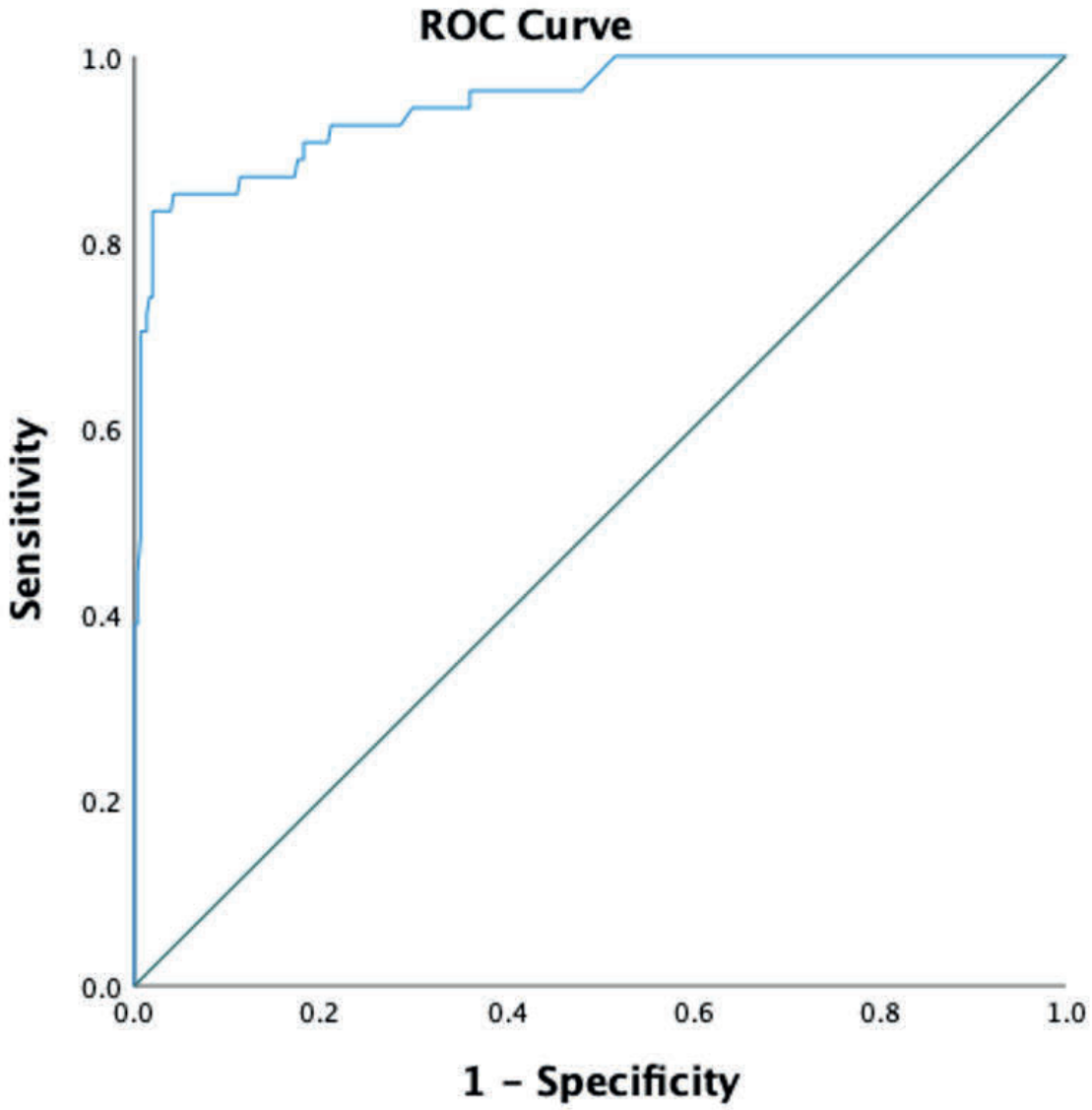
Binary Logistic Regression for Narrative Dynamic Assessment Predictor Variables

Model	Step	Predictor	*Beta*	*Odds Ratio*	R^2	ΔR^2	χ^2	Wald	Sens.	Spec.	AUC
1	1	Posttest	-.17	.84	.40		92.41**	9.08	.83	.76	.87
	2	Mod Judge	-2.59	.08	.70	.30	91.65**	45.65	.89	.83	.95
2	1	Posttest	-.17	.84	.40		92.41**	9.08	.83	.76	.87
	2	Mod Total	-.13	.88	.66	.26	77.50**	1.22	.93	.80	.94
	3	Mod Judge	-.21	.12	.70	.04	15.38**	13.30	.91	.82	.95

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

Figure 1

Receiver Operator Results (ROC) Analysis Yielding Area Under the Curve (AUC) Results With Optimal Cut-Points for Sensitivity and Specificity.



Discussion

The purpose of this study was to examine and cross-validate how well posttest and modifiability scores from the dynamic assessment of language reported in Petersen et al. (2017) account for variance in language ability in a large sample of school-age students and to examine

how well those variables can accurately identify language disorder. Hierarchical logistic regression indicated that the combination of modifiability scores and posttest scores from the dynamic assessment accounted for 70% of the variance in language ability. ROC analyses yielding an AUC using these dynamic assessment variables revealed good to excellent sensitivity and specificity-consistently above 85% in a large sample of students. Given these results, dynamic assessment is a viable option for identifying language disorder for CLD populations.

Additionally, this study addressed weaknesses described in previous dynamic assessment research. For example, in a meta-analysis, Orellana et al. (2019) reported that prior dynamic assessment research (a) failed to identify which variables from dynamic assessment were predictive of language disorder, (b) consistently included small sample sizes – 55 participants or less, (c) lacked a representative ratio of students with and without language disorder, (d) typically did not include a one-gate design, where examiners were blind to language ability prior to testing, and (e) had limited replicability of procedures and cross-validation. This study aimed to address each of these weaknesses.

The current study specifically identified the combination of modifiability scores and posttest scores as significantly predicative dynamic assessment variables for language disorder. These dynamic assessment variables accounted for 70% of the variance. In Petersen et al. (2017), Peña et al. (2006), and Peña et al. (2014) posttest and modifiability variables were also shown to be most predictive of language disorder. Since modifiability scores are sensitive to a student's ability to learn, this study has shown that their predictive power can be greater than results from a static test. A student’s difficulty to learn, specifically to learn language, is what characterizes a language disorder (Bishop, 2017; Bishop et al., 2016). Thus, as this study shoes, modifiability scores more closely reflect the construct of a language disorder, which subsequently aids in

yielding high classification accuracy. Furthermore, modifiability is a measure of a child's ability to learn something new instead of what a student currently can do, which is often confounded by extraneous variables such as English language proficiency, socioeconomic status, and prior education. Lastly, the combination of the student's modifiability scores and their posttest retell scores yield the best sensitivity and specificity. Posttest scores add to the predictive power of modifiability scores because students with a language disorder will not learn as much as those without language disorder after the teaching phase.

This study included the largest sample of participants to date in a dynamic assessment study with a sample size of 362 students. All previous studies had far fewer participants, which potentially impacted the accuracy and generalizability of the results (Orellana et al., 2019). This study's large sample size can provide greater evidence of validity for the dynamic assessment used in the current study and can provide greater confidence in generalizing these results to different populations. Additionally, this study included a sample that had a fairly representative ratio of students with and without language disorder, with 15% having language disorder (Tomblin et al., 1997). Thus, since this study did include a representative ratio in its sample, there is greater confidence in the diagnostic accuracy of the dynamic assessment.

The current study also used a one-gate design where the examiners were blind to language ability prior to and during testing. All studies except for two (Kramer et al., 2009; Peña, 2000) in the Orellana et al. (2019) meta-analysis did not use this design. A one-gate design can increase the evidence of validity of an assessment because examiners who are aware of the diagnosis of students prior to administering a dynamic assessment may score the pretest and posttest and rate the modifiability of students in a biased manner. Therefore, a one-gate design

like the one used in this study can decrease biased scoring and increase evidence for dynamic assessment validity.

Finally, this study carefully details the procedures used to administer the dynamic assessment, which allows for replicability and cross-validation. This study cross-validates and replicates previous dynamic assessment research with similar significant findings, indicating greater confidence in the diagnostic accuracy of the dynamic assessment and its procedures. This study cross-validated findings from Petersen et al. (2017) by administering the same dynamic assessment to an independent, significantly larger sample of students using a relatively large number of examiners. Since 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 that can be generalized to a greater and more diverse 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% with a relatively large and diverse sample.

Clinical Implications

Need for Valid Language Assessments

The majority of traditional NRTs have resulted in poor classification accuracy and there is an increasing need to develop valid and reliable language assessments, particularly with CLD populations (Denman et al., 2017; Friberg, 2010; Spaulding et al., 2006). The results of this study indicate that the dynamic assessment used in this study is a promising alternative to traditional NRTs for diverse groups of students. The sensitivity and specificity of the dynamic assessment and diagnostic results are consistent with other language dynamic assessment studies

for language (e.g., Peña et al., 2006; Peña et al., 2014; Petersen et al. 2017). The research is evident that dynamic assessment has superior classification accuracy over most traditional NRTs, particularly when administered to culturally and linguistically diverse students where differentiation between difference and disorder is difficult. When assessments lack adequate sensitivity and specificity, the chance of misidentifying children with and without disorders increases. It is important to have proper identification methods for all students through valid testing measures, regardless of background. It is common to find disproportionate representations of CLD students in special education at higher or lower percentage rates than other students (Donovan & Cross, 2002). Thus, the results of this study indicate that the dynamic assessment may be a valid tool for proper identification of language disorder across all populations.

Role of Dynamic Assessment in Treatment

The DYMOND can be used clinically as a valid assessment to identify language disorder in diverse populations when considering the modifiability rating scores and/or the posttest. Both the total modifiability score, the final modifiability score and/or the posttest provide vital clinical information for classification and treatment. In contrast to typical NRTs, a dynamic assessment of narrative language can not only provide more accurate classification, but it can provide clinically relevant information for goals and treatment. The teaching phase of the dynamic assessment examined in this study allows the clinician an individualized description of a student's narrative strengths and weaknesses and an opportunity to scaffold support throughout. Specifically, the dynamic assessment used in this study focuses on narrative language which allows for functional goals on narrative discourse and complex language features. Narrative language is replete with academic language and features that are essential for school-based

therapy and a student's classroom success (Petersen, 2011; Spencer et al., 2014). Therefore, this dynamic assessment includes vital diagnostic and clinical data useful for every clinician.

Study Limitations and Future Research

To our knowledge, this was the largest dynamic assessment of language study to date. However, although this study included a relatively large sample of white and Hispanic students, other ethnicities and races are not well represented. This lessens the potential generalizability of the diagnostic accuracy results to other races or ethnicities. Additionally, the sample was exclusively from the Mountain West of the United States, possibly decreasing the generalizability of the dynamic assessment to students in other regions. Future research may need to include more students from other ethnicities and races and from varying locations. Additionally, this study included only 7 Kindergarten students and only 2 of which had language disorder and there were no sixth graders who had language disorder. Thus, sensitivity and specificity results recorded in this study for Kindergarten and sixth grade need to be interpreted with caution. More Kindergarten and sixth grade students need to be included in future research. Additionally, this study used logistic regression to combine the modifiability variables and the posttest score which provided the researchers with a probability variable that was used in the ROC analysis to obtain the AUC and sensitivity and specificity values. The AUC and sensitivity and specificity values allowed for cross-validation and replication of previous studies and to determine the validity of the dynamic assessment. However, future research should identify the cut-points for modifiability and posttest variables that are clinically interpretable.

References

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing words to life: Robust vocabulary instruction*. Guilford.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*(2), 133–146.

Bishop, D. V. M. (2017). Why is it so hard to reach agreement on terminology? The case of developmental language disorder (DLD). *International Journal of Language & Communication Disorders*, *52*(6), 671–680.

Bishop, D. V., Snowling, M. J., Thompson, P. A., Greenhalgh, T., & CATALISE consortium (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PloS one*, *11*(7), e0158753. https://doi.org/10.1371/journal.pone.0158753

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly, 30*(4), 226-236. doi: http://dx.doi.org.erl.lib.byu.edu/10.1177/1525740108326334

Colby, S. & Ortman, J. (2015). *Projections of the size and composition of the U.S. population: 2014 to 2060.* (Current Population Reports; Report number P25-1143). U.S. Census Bureau.https://www.census.gov/content/dam/Census/library/publications/2015/demo/p25-1143.pdf

Denman, D., Speyer, R., Munro, N., Pearce, W. M., Chen, Y. W., & Cordier, R. (2017). Psychometric properties of language assessments for children aged 4-12 years: A

systematic review. *Frontiers in psychology*, *8*, 1515. https://doi.org/10.3389/fpsyg.2017.01515

Dollaghan, C., & Campbell, T. F. (1998). Nonword repetition and child language impairment. *Journal of Speech, Language, and Hearing Research, 41(5),* 1136–1146.

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

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.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy, 26*(1), 77-92. https://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*(2), 103–127. https://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:10.1177/0265659007072142

IBM Corporation. (2016). *IBM SPSS statistics for windows* (Version 24.0) [Computer Software]. IBM Corp. https://www.ibm.com/us-en/

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. 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*(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://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.

Orellana, C. I., Wada, R., & Gillam, R. B., (2019). The use of dynamic assessment for the diagnosis of language disorders in bilingual children: A meta-analysis. *American Journal of Speech-Language Pathology*, *28*(3), 1298-1317

Patterson, J. L., Rodríguez Barbara 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*(1), 103–112. https://doi: 10.1044/1058-0360(2012/11-0129)

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based SLPs: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258.

Peña, E., & Iglesias, A. (1992). The application of dynamic methods to language assessment: A nonbiased procedure. *The Journal of Special Education, 26*(3), 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*(2), 138-154. https://doi:10.1044/1058-0360(2001/014)

Peña, E. D. (2000). Measurement of modifiability in children from culturally and linguistically diverse backgrounds. *Communication Disorders Quarterly, 21*(2), 87–97. https://doi.org/10.1177/152574010002100203

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. 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*(5), 1037-1057. https://doi:10.1044/1092-4388(2006/074)

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

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*, 3–21. https://doi.org/10.1177/0022219413486930

Petersen, D. B., & Spencer, T. D. (2012). The narrative language measures: tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education, 19,* 119-129.

Romero, M. F. (2019). *The accuracy of a Spanish dynamic assessment of narrative language in identifying language disorder: A cross validation study* [Master's thesis, Brigham Young University]. BYU ScholarsArchive Theses & Dissertations. https://scholarsarchive.byu.edu/etd/8271/

Roseberry, C. A., & Connell, P. J. (1991). The use of an invented language rule in the differentiation of normal and language-impaired spanish-speaking children. *Journal of Speech and Hearing Research*, *34*(3), 596–603.

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

Spencer, T.D., Kajian, M., Petersen, D.B., & Bilyk, N. (2013). Effects of an individualized narrative intervention on children's storytelling and comprehension skills. Journal of Early Intervention, 35(3), 243-269.

Tomblin, J. B., Records, N. L., Buckwalter, P., Zhang, X., Smith, E., & O'Brien, M. (1997). Prevalence of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research*, *40*(6), 1245–1260.

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.

Vespa, J., Armstrong, D. M., & Medina, L. (2018). *Demographic turning points for the United States: Population projections for 2020 to 2060*. (Current Population Reports; Report number P25-1144). U.S. Census Bureau.

Westby, C. (1985). Learning to talk—Talking to learn: Oral literate language differences. In C. S. Simon (Ed), *Communication skills and classroom success: Therapy methodologies for language learning disabled students* (pp. 182-213). College Hill.

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. https://doi.org/10.3109/17549507.2011.636071

APPENDIX A

Annotated Bibliography

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools*, *44*(2), 133–146.

Objective: The purpose of this study was to investigate if the classification quality of NRTs effect the amount they are used clinically by SLPs. *Methods:* The researchers surveyed 364 SLPs from across the United States through an online format. 55 NRTs for language were included and analyzed for diagnostic accuracy. *Results:* SLPs do not take diagnostic accuracy into account when choosing and administering a test. Some NRTs did report the sensitivity and specificity. The CELF-4 and the PLS-5 are the most commonly used based on year of publication. *Relevance to work:* There is a need for increased evidence based practice and assessments with higher diagnostic accuracy.

Caesar, L. G., & Kohler, P. D. (2009). Tools clinicians use: A survey of language assessment procedures used by school-based speech-language pathologists. *Communication Disorders Quarterly, 30*(4), 226-236. doi: http://dx.doi.org.erl.lib.byu.edu/10.1177/1525740108326334

Objective: This study aimed to gather information regarding assessment tools and specific procedures used by school- based SLPs. *Methods:* A mail in survey was sent to 409 school-based SLPs in the state of Michigan. The questionnaire included demographic information as well as informal versus formal assessment use. *Results:* The findings showed that most SLPS use the combination of norm referenced tests (NRTs) like the CELF and informal assessments like parent reports when assessing students. Dynamic assessment was not used

by most SLPs and assessment procedures often did not change for CLD students. *Relevance to work:* School-based SLPs are more likely to use NRT's for assessment for all students including CLD populations and are not often utilizing dynamic assessment.

Friberg, J. C. (2010). Considerations for test selection: How do validity and reliability impact diagnostic decisions? *Child Language Teaching and Therapy,26*(1), 77-92. doi:10.1177/0265659009349972

Objective: The purpose of this article is to assist clinicians when considering a standardized assessment by applying psychometric criteria discussed for nine specific tests and a decision tree when considering other assessments. *Methods:* The article evaluated 9 preschool and school-aged standardized language assessments based on an identification accuracy of .80 or better and availability. The tests included CELF-4, CELF-P2, PLS-4, SPELT-3. SPELT-P2, TEEM, TEGI, TLC-E and TNL (see page 80 (4) of pdf). These tests were examined with 11 specific psychometric criteria to determine validity: (a) purpose of the test was identified, (b) test qualifications are explicitly stated, (c) testing procedures are sufficiently explained, (d) adequate standardization sample size (>100) is noted, (e) clearly defined standardization sample including information related to geographic representation, SES, parent education, gender distribution, ethnic background, presence/absence of impairment and age, (f) evidence of item analysis, (g) measures of central tendency are reported, 8_ concurrent validity is documented, (h) predictive validity is documented, (i) test/re-test reliability is reported, and (j) inter-rater reliability is reported. *Results:* The evaluation concluded that all assessments met at least 8 out of 11 criteria and are considered to have acceptable classification accuracy levels. However, no assessments met all 11 and the tests were found to be in a range of 8-10

criteria per test. Of the 9 tests, 78% did not meet the predictive validity criteria and 44% did not meet the test-retest criteria. Only the TNL (Gillam and Pearson, 2004) and the TEEM (Shipley et al., 1983) reported predictive validity in their tests. *Relevance to current work:* Out of all the possible standardized assessments, only 9 were considered acceptable. These tests also lacked predictive validity and representative samples. Thus, assessments with high classification accuracy for CLD populations are needed.

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: This study examined the classification accuracy of the Dynamic Assessment and Intervention (DAI) tool created by Miller et al. in 2001. *Methods:* 17 third grade children from a First Nations community were administered the DAI tool. *Results:* The normal language learning group and the possible language learning difficulty group benefitted from the teaching phase. The participants in the normal language learning group, however, demonstrated greater gains and generalization of targets that were not specifically addressed in the teaching phase. Discriminant analyses revealed high sensitivity and specificity. *Relevance to work:* Dynamic assessment in general can be a useful tool is accurately diagnosing language disorders.

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

Objective: This clinical forum discusses problems with norm-referenced testing and recent solutions that have been developed to reduce bias in assessments for growing

culturally and linguistically diverse (CLD) populations. *Methods:* The forum delineated common limitations with norm-referenced testing like content bias, linguistic bias, and disproportionate representation in normative samples. It also discussed criterion-referenced measures, processing-dependent measures and the use of dynamic assessments for CLD populations. It outlined specific methods for dynamic assessments that include test-teach-retest models, task/stimulus variability models and graduated prompting models. *Relevance to current work*: A promising alternative to biased norm-referenced testing are dynamic assessments. They are based on a child's zone of proximal development and can determine current level of functioning and the best way to facilitate learning for the individual child. This diagnostic tool can reduce bias in CLD populations and are quick and easy to administer.

Orellana, C. I., Wada, R., & Gillam, R. B. (2019). The use of dynamic assessment for the diagnosis of language disorders in bilingual children: A meta-analysis. *American Journal of Speech-Language Pathology*, 28, 1298-1317

Objective: The purpose of this study was to provide a systematic review of the diagnostic accuracy of dynamic assessments for language impairment in bilingual children and to examine their current clinical use. *Methods:* Seven studies were reviewed using a meta-analysis procedure outlined by Cooper (2017) that included study identification, inclusion criteria, search results, coding procedure and interrater reliability. The studies included participants with a range of 3-8 years old and varying language areas including, labeling single words, morpheme rule learning, ability to learn nonwords, and narratives. *Results:* The dynamic assessment studies demonstrated higher scores on language assessments for typically developing (TD) participants than those with language impairment.

Additionally, modifiability scores during the teaching phase consistently yielded higher scores for TD participants. However, gain scores from the pretest-teach-posttest models were less likely to accurately identify language impairment. The researchers concluded that using dynamic assessment with static assessments, case history, questionnaires, and observations of the child may lend to more accurate assessments and identification of language impairment in bilingual children. *Relevance to current work:* Dynamic assessments yield good classification accuracy for bilingual children especially when using a clinician's judgement of modifiability during a teaching phase. Specifically, Petersen's (2017) dynamic assessments of narratives received 7/9 quality indicators for diagnostic accuracy.

Pavelko, S. L., Owens, R. E., Jr., Ireland, M., & Hahs-Vaughn, D. L. (2016). Use of Language Sample Analysis by School-Based SLPs: Results of a Nationwide Survey. *Language, Speech, and Hearing Services in Schools*, *47*(3), 246–258.

Objective: This article examined the use of language sample analysis (LSA) by speech language pathologists (SLPs) in school-based assessments to further explore the practice of LSA use in public schools across the United States, determine the characteristics found in LSAs when they were used and how SLPs transcribed and analyzed these samples. *Methods:* 1,399 school-based SLPs from 34 different states responded to a 28-question survey across three general areas: use of LSA in clinical practice, use of standardized testing in clinical testing, and training needs. *Results:* A total of 893 (67%) of all respondents reported using LSA during the 2012–2013 school year. Of those responding *yes* to the use of LSA, more than half analyzed 10 or fewer language samples (55%). Of the respondents who reported they did not use LSA, 66% reported that language sampling is a

requirement for eligibility for services in their state. SLPs working in middle school and high school were less likely to use LSA in their assessments. Lastly, 78% of respondents reported that they did not use LSA because it is too time-consuming. *Relevance to current work:* Important findings in this article include the number of SLPs using LSA to assess and qualify school-age children for services, in spite of evidence-based practices relative to LSA (Lund & Duchan, 1993; Nippold, 2014). Additionally, previous research has reported a notable number of SLPs not using LSA (Kemp & Klee, 1997), however, the current study found an even higher percentage.

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: This study examined the reliability and classification accuracy of a narrative dynamic assessment for school-age students. *Methods:* First, 58 first and second grade students narrated wordless picture books, then 71 students retold narratives using dynamic assessment procedures (pretest, teach, posttest). *Results:* The researchers found that students who were typically developing saw a greater change from their pretest to posttest scores than the students with language disorder. Sensitivity and specificity values were highest when modifiability scores and posttest scores were used. *Relevance to work:* Dynamic assessments of narrative language may be a promising alternative to NRTs. They also have high classification accuracy and high sensitivity and specificity when modifiability sores and posttest scores are used.

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:* The purpose of this study was to identify the classification accuracy of an English dynamic assessment of narrative language in identifying language disorders. *Methods:* This study consisted of 42 Spanish-English bilingual K-3 children. Each participant was administered two 25-miuntes dynamic assessments with a pretest-teach-posttest model. The pretest and posttest consisted of the retelling of narratives and were scored in real time. During the teaching phase, examiners taught children missing story grammar elements and language complexity targets. *Results:* A 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 dynamic assessment session and 100% sensitivity and specificity after the second dynamic assessment 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 dynamic assessment 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 the modifiability scores and posttest yield high sensitivity and specificity. Gain scores may not be not be very predictive or helpful. The dynamic assessment can be further abbreviated and maintain high classification accuracy.

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

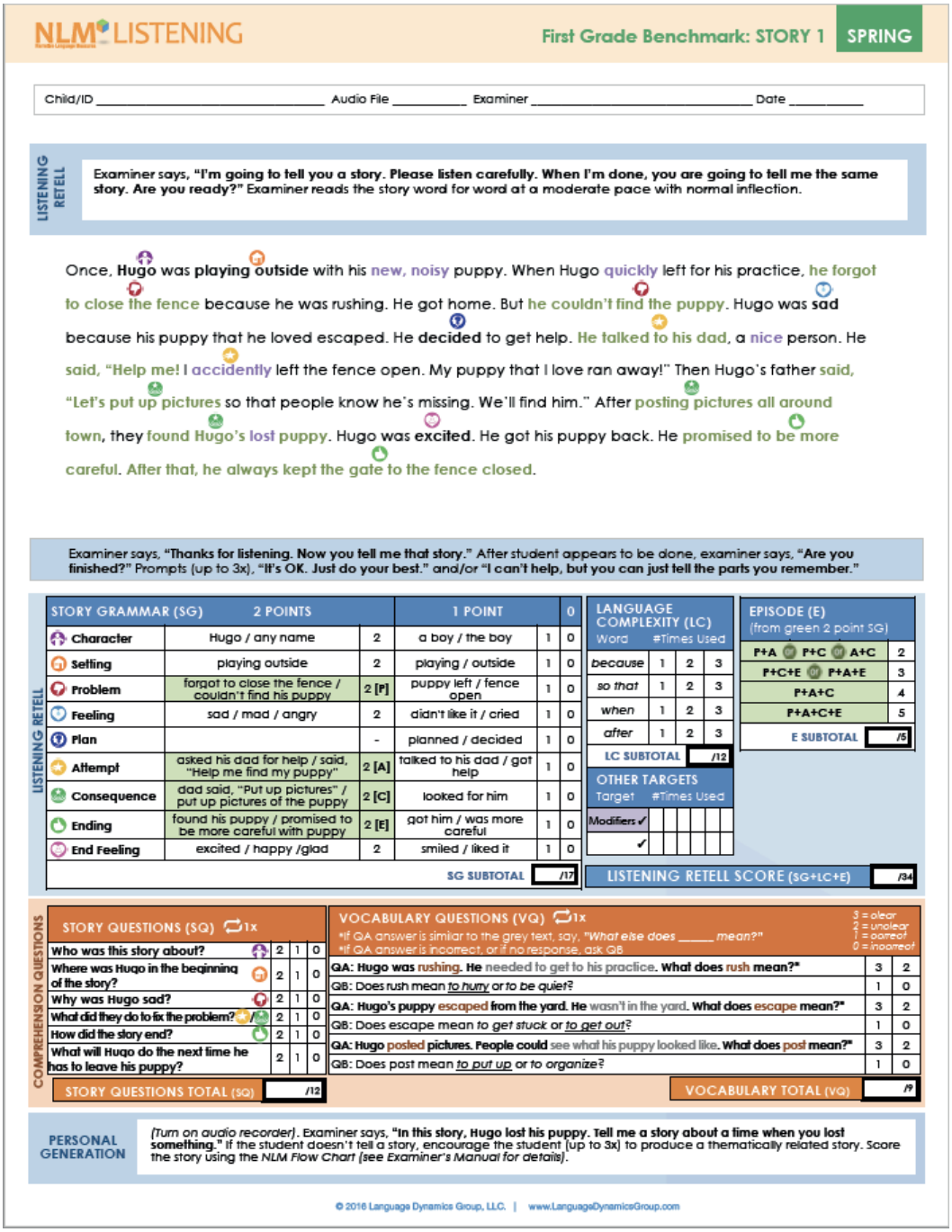
Objective: This article examined the classification accuracy of standardized assessments for language impairment based on the use of low scores and data in the manuals of tests to identify a disorder. *Methods:* The latest edition of 43 commercially available norm-referenced standardized language assessments for children ages 3 to 18 years were reviewed by 3 clinically certified and experienced SLPs. Data included mean differences in subtest scores, test composites scores, and/or total test scores in samples of children with language impairment and specific information concerning the sensitivity and specificity of each test. *Results:* The data collected failed to prove that low performance scores on a test's normative distribution qualifies a child with a language impairment. The mean group differences found in these tests suggest that children with language impairment are likely to score closer to the normative sample as opposed to a cut-off score. Additionally, sensitivity and specificity ratings were only available in 9 out of 43 tests and 5 of those reported an acceptable accuracy of 80% or better. *Relevance to current work:* This article emphasized that an evidence-based practice framework for diagnostic accuracy is measured primarily through sensitivity and specificity data. Currently, most commonly norm-referenced tests lack adequate sensitivity and specificity for proper identification and can result in high numbers of misidentification. Thus, assessments with adequate sensitivity and specificity are needed.

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: This study examined the teaching of phonemic awareness and the classification accuracy of this dynamic word learning assessment. *Methods:* 36 children from ages 5:0 to 6;5 years old participated in a 7-week intervention process in small groups. The dynamic assessment included four phonemic awareness tasks. *Results:* The study found that the typically developing participants scored higher in phonemic awareness assessments than the group with language impairment. *Relevance to work:* A dynamic assessment of language can accurately identify students with and without language disorder.

APPENDIX B

First-Grade Spring Benchmark Story 1 Used to Help Identify Language Ability



NLM LISTENING — First Grade Benchmark: STORY 1 — SPRING

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.

Once, **Hugo** was **playing outside** with his new, noisy puppy. When Hugo quickly left for his practice, he forgot to close the fence because he was rushing. He got home. But he couldn't find the puppy. Hugo was **sad** because his puppy that he loved escaped. He **decided** to get help. He talked to his dad, a nice person. He said, "Help me! I accidently left the fence open. My puppy that I love ran away!" Then Hugo's father said, "Let's put up pictures so that people know he's missing. We'll find him." After posting pictures all around town, they found Hugo's lost puppy. Hugo was **excited**. He got his puppy back. He promised to be more careful. After that, he always kept the gate to the fence closed.

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
Character	Hugo / any name	2	a boy / the boy	1	0
Setting	playing outside	2	playing / outside	1	0
Problem	forgot to close the fence / couldn't find his puppy	2 [P]	puppy left / fence open	1	0
Feeling	sad / mad / angry	2	didn't like it / cried	1	0
Plan		-	planned / decided	1	0
Attempt	asked his dad for help / said, "Help me find my puppy"	2 [A]	talked to his dad / got help	1	0
Consequence	dad said, "Put up pictures" / put up pictures of the puppy	2 [C]	looked for him	1	0
Ending	found his puppy / promised to be more careful with puppy	2 [E]	got him / was more careful	1	0
End Feeling	excited / happy /glad	2	smiled / liked it	1	0
			SG SUBTOTAL		/17

LANGUAGE COMPLEXITY (LC) Word	#Times Used		
because	1	2	3
so that	1	2	3
when	1	2	3
after	1	2	3
LC SUBTOTAL			/12

OTHER TARGETS Target	#Times Used
Modifiers ✓	
✓	

EPISODE (E) (from green 2 point SG)	
P+A or P+C or A+C	2
P+C+E or P+A+E	3
P+A+C	4
P+A+C+E	5
E SUBTOTAL	/5

LISTENING RETELL SCORE (SG+LC+E) /34

COMPREHENSION QUESTIONS

STORY QUESTIONS (SQ) 1x			
Who was this story about?	2	1	0
Where was Hugo in the beginning of the story?	2	1	0
Why was Hugo sad?	2	1	0
What did they do to fix the problem?	2	1	0
How did the story end?	2	1	0
What will Hugo do the next time he has to leave his puppy?	2	1	0
STORY QUESTIONS TOTAL (SQ)			/12

VOCABULARY QUESTIONS (VQ) 1x
*If QA answer is similar to the grey text, say, "What else does _____ mean?"
*If QA answer is incorrect, or if no response, ask QB
3 = clear, 2 = unclear, 1 = correct, 0 = incorrect

QA: Hugo was rushing. He needed to get to his practice. What does rush mean?"	3	2
QB: Does rush mean to hurry or to be quiet?	1	0
QA: Hugo's puppy escaped from the yard. He wasn't in the yard. What does escape mean?"	3	2
QB: Does escape mean to get stuck or to get out?	1	0
QA: Hugo posted pictures. People could see what his puppy looked like. What does post mean?"	3	2
QB: Does post mean to put up or to organize?	1	0
VOCABULARY TOTAL (VQ)		/9

PERSONAL GENERATION

(Turn on audio recorder). Examiner says, **"In this story, Hugo lost his puppy. Tell me a story about a time when you lost something."** 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

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