Brigham Young University
BYU ScholarsArchive

Theses and Dissertations

2020-08-06

The Classification Accuracy of a Dynamic Assessment of Inferential Word Learning for School-Age Children With and Without Language Disorder

Britney Ann Newey
Brigham Young University

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

 Part of the Communication Sciences and Disorders Commons

BYU ScholarsArchive Citation

Newey, Britney Ann, "The Classification Accuracy of a Dynamic Assessment of Inferential Word Learning for School-Age Children With and Without Language Disorder" (2020). *Theses and Dissertations*. 8672.
https://scholarsarchive.byu.edu/etd/8672

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

The Classification Accuracy of a Dynamic Assessment of Inferential Word Learning for School-Age Children With and Without Language Disorder

Britney Ann Newey

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

Master of Science

Douglas B. Petersen, Chair
Shawn L. Nissen
Kathryn L. Cabbage

Department of Communication Disorders

Brigham Young University

Copyright © 2020 Britney Ann Newey

All Rights Reserved

ABSTRACT

The Classification Accuracy of a Dynamic Assessment of Inferential Word Learning for School-Age Children With and Without Language Disorder

Britney Ann Newey
Department of Communication Disorders, BYU
Master of Science

Purpose: This study examines the classification accuracy and interrater reliability of a dynamic assessment (DA) of inferential word learning designed to accurately identify kindergarten through sixth-grade students with and without language disorder. Method: The participants included 127 school-age children from a mountain west school district who were administered a DA of inferential word learning that entailed a pretest, a teaching phase, an examiner rating of the child's ability to infer word meaning (modifiability), and posttests. Results: Hierarchical logistic regression and receiver operator characteristic (ROC) analyses revealed that combining all posttests, the modifiability total, and the final examiner judgement scores from this DA yielded the strongest sensitivity (.83) and specificity (.80). The static measures and the dichotomized final examiner judgement had excellent reliability; yet the individual modifiability measures (with the exception of disruption and frustration) had poor reliability. Conclusion: In concordance with a previous study, results indicate that a dynamic assessment of inferential word learning may be an efficacious method of identifying language disorders in school-age populations.

Keywords: dynamic assessment, context clues, vocabulary, inferential word learning

ACKNOWLEDGMENTS

I would like to thank the following committee members for their expertise and guidance: Dr. Douglas B. Petersen, Dr. Shawn L. Nissen and Dr. Kathryn L. Cabbage. I would especially like to thank Dr. Petersen for his encouragement, flexibility, and patience. I would also like to thank Lisa Wells for her insightful assistance during the editing process.

TABLE OF CONTENTS

LIST OF TABLES

DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *The Classification Accuracy of a Dynamic Assessment of Inferential Word Learning for School-Age Children With and Without Language Disorder*, is written in a hybrid format to adhere to traditional thesis requirements and journal publication formats. The initial pages of the thesis adhere to university requirements while the thesis report is presented in journal article format. The Dynamic Assessment of Inferential Word Learning (DA-IWL) Scoring Sheet and annotated bibliography are included in appendices.

Introduction

Speech-Language Pathologists (SLP) frequently use norm-referenced vocabulary assessments when diagnosing school-age children with language disorders (LD). However, most static norm-referenced vocabulary tests do not identify LD accurately, having low sensitivity and specificity. For example, Gray, Plante, Vance, and Henrichsen (1999), investigated the sensitivity and specificity of the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), Receptive One-Word Vocabulary Test (Gardner, 1985), Expressive Vocabulary Test (Williams, 1997), and Expressive One-Word Vocabulary Test-Revised (Gardner, 1990). The researchers administered the tests to 62 preschool-age children—31 with an LD and 31 with Typically developing language (TD). Approximately half (15/31 LD, 17/31 TD) of the participants were misclassified by at least one of the tests, illustrating the tests' low sensitivity (71% - 77%) and specificity (68%-77%). Additionally, in their 2006 review of 43 norm-referenced language assessments, Spaulding, Plante, and Farinella (2006) found that vocabulary assessments in particular had poor classification accuracy. One possible explanation for this phenomenon is that static vocabulary assessments measure children's current vocabulary knowledge and do not control for differences in language experiences or dialectal differences. Vocabulary development is highly dependent upon a child's individual language experiences. These factors often result in misidentification of children from culturally and linguistically diverse (CLD) backgrounds.

Word knowledge is an important aspect of language development and grows at an accelerated rate in young children. At five years-of-age, a typically developing English-speaking child has a vocabulary of 4,000 to 5,000 words (Nation & Waring, 1997) and once formal schooling begins, children learn approximately 2,000 to 4,000 words each year (Baumann & Kameenui,1991). The specific words in a child's vocabulary are highly dependent on the child's

experiences and consequent exposure to certain words. Most of the vocabulary that children learn is acquired through inferential word learning. Inferential word learning is the process of deducing the meaning of an unknown word through use of context clues (Petersen, Tonn, Spencer, & Foster, 2019). However, children with LD commonly have much smaller lexicons (Nation & Waring, 1997) and more difficulty acquiring new words. Because of this, *vocabulary learning* could be a strong diagnostic marker for LD. An assessment that measures inferential word learning may have stronger sensitivity and specificity over a static measure that only assesses what a child currently knows.

Dynamic Assessment

Dynamic assessment (DA) is an application of cognitive psychologist Reuven Feuerstein's theory that intelligence is fluid, rather than static. He believed that educators can help children develop critical thinking skills through mediated learning experience (teaching) and that a child's learning potential can be measured through DA. Thus, DA aims to measure an individual's potential to change, or their modifiability. It measures a child's ability to learn rather than their current knowledge (Feuerstein, Rand, & Hoffman, 1979).

There are two main approaches to DA that have strong evidence of validity (Orellana, Wada, & Gillam, 2019): graduated prompting and the test-teach-retest model. The graduated prompting model attempts to determine the amount of adult support a child needs to be successful in a given learning task. The examiner provides the child with various hierarchical prompts which range from indirect to more directive prompts. In this model, the less adult support a child needs, the more modifiable the child is considered. Another type of DA consists of three phases: a pretest, a mediated learning experience, and a posttest. The pretest gives the examiner a measurement of the child's current ability to complete the task. During the mediated

learning experience or teaching phase, the examiner teaches the child the desired skill and targets the pretest deficits. Following the teaching phase, the examiner rates the child's learning skills (e.g., self-regulation and attention) and the amount of examiner effort required to teach the child. These ratings are typically combined into an overall modifiability index. The examiner then retests the child to determine the amount of information the child independently transfers from the teaching phase.

Research on Dynamic Assessment and Vocabulary

Various researchers in the field of speech-language pathology have investigated the use of DA of vocabulary in the diagnosis of LDs. For example, Peña, Quinn, and Iglesias (1992) illustrated the validity of a test-teach-retest DA in the context of vocabulary labeling strategies. When they combined posttest and modifiability scores, they were able to correctly identify the language abilities of 93% of the children. In similar studies, Gutiérrez-Clellen and Peña (2001), Peña, Iglesias, and Lidz (2001), Ukrainetz, Harpell, Walsh, and Coyle (2000), and Kapantzoglou, Restrepo, and Thompson (2011) found that modifiability scores from vocabulary DAs strongly differentiated between children with strong and weak language skills. Gutiérrez-Clellen and Peña found that a test-teach-retest method of DA is particularly effective in differentiating between LDs and language differences. They also found that posttest and modifiability scores provide the most diagnostically valuable information and Peña, Iglesias, and Lidz further confirmed this finding. Rather than targeting specific vocabulary words, Ukrainetz et al. targeted vocabulary categorization in a DA administered to Native-American children with strong (N = 15) and weak (N = 8) language abilities. The response to mediation checklist (a modifiability measure of transfer and responsivity) separated strong and weak language learners with high accuracy (87% sensitivity and 100% specificity). In 2011, Kapantzoglou et al. looked

at a DA of word learning skills in English-Spanish bilingual children. They used a scripted, structured play activity as their teaching phase. All children were classified with 76.9% sensitivity and 80% specificity. Again, the modifiability score was the strongest separating factor between groups.

Camilleri and Law (2007), Camilleri and Botting (2013), and Camilleri and Law (2014) investigated a DA of receptive vocabulary skills using a test-teach-retest design. Camilleri and Law (2007) compared the ability of DA of word learning to classify bilingual children with typical and atypical language skills to that of a static vocabulary measure, the British Picture Vocabulary Scale II (BPVS; Dunn, Dunn, Whetton, & Burley, 1997). On the static measure, the monolingual English-speaking children scored significantly higher than their multilingual peers. However, on the dynamic measure, the gap between the bilingual and monolingual groups narrowed, revealing no significant differences. In 2013, Camilleri and Botting developed this testing framework into the Dynamic Assessment of Word Learning (DAWL). They investigated the predictive validity of the DAWL after six months of the initial test administration. They found the scores accurately predicted the vocabulary skills of children with low language ability. In 2014, Camilleri and Law compared the predictive validity of the BPVS to that of the DAWL for both children with high and low language abilities. They found the DAWL was more predictively relevant for children with low language abilities.

Larsen and Nippold (2007), Ram, Marinellie, Benigno, and McCarthy (2013), and Wolter and Pike (2015) investigated the use of a graduated prompting DA approach to assessing word learning through morphological analysis in school-age students. Larsen and Nippold conducted a preliminary study of the Dynamic Assessment Task of Morphological Awareness (DATMA) with 50 typically developing sixth-grade monolingual English speakers. In this version of the

DATMA, students were expected to define 15 target words. These words consisted of low frequency derivations of high frequency words. This study showed that the DATMA appropriately assessed morphological awareness in typically developing sixth graders. In a replication study, Ram et al. administered a modified DATMA to both third and sixth graders and investigated the effect of presenting target words in isolation vs. context. Children needed fewer prompts to determine word meaning when target words were presented in context. Additionally, third graders required context clues more often than sixth graders when defining words. Wolter and Pike also found that DA is a valuable tool for measuring early derivational morphological awareness skills in typically developing third graders. These young, typically developing students required high amounts of adult scaffolding to identify word meaning using morphological cues. This suggests that measuring morphological awareness in young children, even when there is considerable adult support, may be problematic. However, children's ability to infer the meaning of words in context may be more appropriate for children in low grades, since children typically develop this skill earlier than morphological awareness skills. Additionally, the researchers hypothesized that a measure of contextualized word-learning may result in higher interrater reliability in lower grades because coding the correctness of students' responses is easier and more straightforward than scoring morphological analysis.

Most recently, Petersen et al. (2019) investigated the ability of a hybrid test-teach-retest and graduated prompting DA of inferential word learning to differentiate between LD and language difference in Spanish/English, bilingual, school-age children. Petersen et al. administered the DA and the Expressive/Receptive Expressive One Word Vocabulary Test–Spanish Bilingual Edition (E/ROWPVT-SBE; Martin & Brownell, 2012a, 2012b), and a standardized narrative retell task using the wordless picture book *Frog Where Are You?* (Mayer,

1969; Miller & Iglesias, 2008) to 31 participants. The DA had much higher sensitivity (90%) and specificity (90.5%) than that of the static assessments. This preliminary study suggests that their DA of inferential word learning has strong classification accuracy and warrants further investigation. Therefore, the purpose of this study is to build on the research conducted by Petersen et al. and determine if a modification of their DA of inferential word learning will identify a larger sample of kindergarten-fifth grade students with and without LDs with adequate sensitivity and specificity, and to determine if this new test has adequate reliability. Sensitivity in this study refers to the percent of children with an LD who were accurately identified as having an LD. Specificity refers to the percent of typically developing children identified as typically developing. The research questions are as follows:

1. What is the sensitivity and specificity of the posttest variables from a dynamic assessment of inferential word learning for kindergarten to fifth-grade students with and without language disorder, and what is the sensitivity and specificity when modifiability variables are added to the posttest variables?
2. What is the interrater reliability of a dynamic assessment of inferential word learning?

Method

Participants

Brigham Young University's Institutional Review Board approved this study. We recruited a total of 127 school-age (K-5) children from general education classrooms and the SLPs' caseloads in a mountain west school district. Additionally, 24% (n = 30) of participants had a language disorder. Table 1 breaks down demographic information by gender and grade level.

Table 1

Demographic Information for All Participants

Demographic type	Total n (%)	Typical language n (%)	Language disorder n (%)
Number of Students	127	97 (76%)	30 (24%)
Gender			
Male	74 (58%)	56 (76%)	18 (24%)
Female	53 (42%)	41(77%)	12 (23%)
Grade Level			
K	5 (4%)	2 (40%)	3 (60%)
1	5 (4%)	1 (20%)	4 (80%)
2	25 (20%)	22 (88%)	3 (12%)
3	16 (13%)	13 (81%)	3 (19%)
4	40 (32%)	29 (72%)	11 (28%)
5	36 (28%)	30 (83%)	6 (17%)

The Index Reference for Language Disorder

We initially classified children as having an LD for our index reference using two different approaches. First, children were classified as having an LD if they had an Individualized Education Program (IEP) and also had low scores on at least one of three indicators from a dynamic assessment of narrative language (total modifiability score of 21 or lower, a final examiner judgment of 1 or lower, and/or a posttest score of 9 or lower) or if they had 70% or less accurate syllables in a nonword repetition task (raw score of 36 or lower). Second, children were classified as having an LD if they did not have an IEP, yet met two of the three following criteria: (a) low scores on at least one of three indicators from a dynamic assessment of narrative language (total modifiability score of 21 or lower, a final examiner judgment of 1 or lower, and/or a posttest score of 9 or lower), (b) 70% or less accurate syllables in a nonword repetition task (raw score of 36 or lower), and/or (c) performance 1.5 standard deviations or lower on a narrative language task (using the Narrative Language Measures; NLM, raw cutoff scores were as follows: kindergarten < 4, first grade < 9, second grade < 14, third

grade < 13, fourth grade < 3, fifth grade < 9). We classified children as having typically developing language when they did not have an IEP and did not meet any of the other criteria outlined for LD.

Measures

We assessed all participants 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.

We used the Narrative Language Measures (NLM) subtest of the CUBED (CUBED; Petersen & Spencer, 2012),6a modified nonword repetition task (Gathercole, Willis, Emslie, & Baddeley, 1992), and a dynamic assessment of inferential word learning to initially evaluate students. The NLM was administered first, followed by the administration of the nonword repetition task and the dynamic assessment. Examiners recorded audio for each of the assessment measures. The examiners were blinded as to whether the students had an LD or not.

CUBED: Narrative Language Measures (NLM). The NLM Listening subtest of the CUBED (Petersen & Spencer, 2016) was used to determine language proficiency and LD. 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 in English. The examiner read the model story to the child and then asked the child to retell the story. The examiner scored and audio-recorded each child's retell in real-time.

Non-word repetition task. We also used a nonword repetition task to help determine whether a child had an LD. We administered 10 nonwords from the Children's Test of Nonword Repetition (Gathercole et al., 1992 and two researcher-generated nonwords. Examiners instructed each student to listen to an audio-recording of the non-words and repeat each word. Later, examiners listened to audio recordings of the students' responses and recorded the number of correct syllables the students' produced and converted this score into a percentage.

Dynamic Assessment of Inferential Word Learning (DA-IWL). The DA-IWL entails several pretests, teaching phases, posttests and a modifiability rating scale, with a greater degree of teaching provided to students who have difficulty on the pretest sections. The dynamic assessment took approximately 10 minutes but varied in length based on student responsiveness.

The pretests consist of short narratives which contain target nonword verbs twice. The examiner begins by saying "I'm going to tell you a short story. Please listen carefully. There is a new word in this story. When I'm done, I'm going to ask you about the new word." After reading the story word for word at a moderate pace, the examiner asks the child to define the nonsense word: "What does [nonsense word] mean"? If appropriate, the examiner waits 5-10 seconds to provide an additional prompt: "It's OK. You can guess." The examiner scores the child's answer on a scale of 0-2 with a total of two possible points on the pretest story. The test protocol included appropriate responses for each stimulus, ensuring scoring reliability. If the student provides a clear, complete definition then the examiner reads a reinforcing script (a simplified teaching phase) which explicitly models strategies used to deduce word meaning and then moves to story two. When children provide an unclear, incomplete, or incorrect definition the examiner moves to teaching phase one, which is more intensive than the reinforcing script. The examiner explains that when they hear an unfamiliar word, they listen to the words before

and after the word to find clues. The examiner then summarizes the clues and asks the child to define the nonsense word as a posttest measure. The examiner again scores the child's answer on a scale of 0-2. If the child still produces an inadequate definition, the examiner defines the word and reads a sentence from the story using the real word in place of the nonsense word. This is an additional prompt provided for those students who cannot infer the meaning of the nonsense word at posttest.

Next the examiner begins pretest two by reading the second story to the child and asking the child to define the nonsense word. The examiner again scores the child's pretest performance on a scale of 0-2. When children score a 2 on this pretest, the examiner proceeds immediately to the modifiability rating. Children who score a 0 or 1 on the second pretest proceed to teaching phase two. This teaching phase uses similar wording and techniques as teaching script one. Again, the child receives another opportunity to define the nonsense word and receives a rating on a scale of 0-2 on the posttest. Immediately after administering the second teaching script, the examiner completes the modifiability rating scale.

The first six questions of the modifiability ratings use a five-point rating scale to describe learning behaviors that the student displays during the test. On the final item, the examiner rates the "student's ability to learn vocabulary through listening to stories" using a five-point rating scale. The examiner then calculates two modifiability scores—a total modifiability index and a final examiner judgement score.

After the modifiability ratings, the examiner administers the posttest story. On the posttest story, the examiner again reads a short story containing a nonsense word and rates the child's definition on a scale of 0-2. Next, the examiner reads a sentence containing a nonsense word and rates the child's definition on a scale of 0-2.

Our DA strictly measured children's verbal ability to learn vocabulary through the use of context clues. During the previous study, examiners asked students to point to one of four pictures that illustrated the meaning of the new word when students were unable to provide an adequate definition.

Fidelity

Before data collection, we trained a team of undergraduate students in the Communication Disorders program at Brigham Young University to administer the assessments. Each team member demonstrated competence and ability to adhere and carry out the testing procedures correctly, consistently, and independently. We extensively trained our team leads who then certified each of our research assistants after an hour-long training. When administering the DA to participants, a team lead was always onsite to monitor adherence to testing procedures.

Results

Data Analysis

We used the Statistical Package for Social Sciences (SPSS version 24.0; IBM Corp., 2016) to analyze data. We conducted logistic regression and receiver operator characteristic (ROC) analyses to determine the sensitivity and specificity of the posttest variables (Model 1) and combined posttest and modifiability variables (Model 2) from the dynamic assessment (Question 1).

Logistic regression utilizes independent and continuous predictor variables to predict a binary dependent variable. In this study, language ability is the binary dependent variable (i.e., LD/no language disorder) and the predictor variables are the inferential word learning dynamic assessment posttest scores and modifiability scores, based on previous DA research findings. We

conducted ROC analyses, which provided area under the curve (AUC) results which we then used to determine the optimal sensitivity and specificity of the dynamic assessment. The AUC provides sensitivity and specificity for each possible cut point of the predictor measure.

Sensitivity and Specificity

We used hierarchical logistic regression to determine to what extent DA variables (posttest 1, 2, and 3, and total modifiability and modifiability final judgment scores) accounted for the variance in language ability (Question 1). In the first hierarchical logistic regression model, we first entered posttest 1 (Step 1, Model 1), then we combined posttest 1 and posttest 2 (Step 2, Model 1), then we combined all three posttests (Step 3, Model 1). In Model 2, we added the total modifiability score to the combined posttests (Step 1, Model 2) and then added modifiability final judgment scores to all three posttests and the modifiability judgment variables (Step 2, Model 2). Model 1 results indicated that all three posttests accounted for 27% of the variance (Nagelkerke $R^2 = .27$) with 72% sensitivity and 74% specificity, and that the combination of the posttest and the modifiability variables accounted for 47% of the variance in language ability (Nagelkerke $R^2 = .47$) with 83% sensitivity and 80% specificity.

We conducted ROC analyses, which provided AUC results, to determine the optimal sensitivity and specificity of the dynamic assessment (Question 2). The AUC provides sensitivity and specificity for each possible cut point of the predictor measure. The overall Wilks's lambda was significant for each of the individual dynamic assessment variables. Table 2 lists Logistic regression and ROC analyses results for each model.

Table 2

Hierarchical Logistic Regression and ROC Analyses with AUC Results for the Posttest and Modifiability DA Variables

Model	Step	Predictor	Beta	*Exp* (B)	R^2	ΔR^2	χ^2	Wald	Sens.	Spec.	AUC
1	1	Posttest 1	-.95	.39	.19	-	17.30**	16.28	.75 (81)	.63 (60)	.71
	2	Posttests 1+2	-.56	.57	.23	.04	21.32**	4.09	.70	.70	.74
	3	All Posttests 1+2+3	-.42	.66	.27	.04	24.86**	2.37	.72	.74	.76
2	1	All Posttests+Mod Total	-.35	.71	.45	.18	43.87**	13.44	.83 (.86)	.79 (.73)	.83
	2	All Posttests+Mod Total +Final Judge	-.74	.48	.47	.02	46.39**	2.31	.83	.80	.82

Note. DA = dynamic assessment. Mod Total = DA modifiability total score. Final Judge = final examiner judgement. AUC = area under the curve. Sens = sensitivity. Spec = specificity. Sensitivity and specificity values in parentheses are alternative results. Beta, Wald, Exp (B) (odds ratio) are from each step of the model. χ^2 degrees of freedom are equal to the number of predictors in each model. $^{**}p < .01$

Test Administration Reliability

We randomly selected approximately 20% (N = 25) of the students' tests from the pool of students with and without LD to calculate inter-rater reliability. We listened to audio recordings of each selected test and scored them without knowledge of the students' language status. We then compared our scores to the original scores. We divided the number of items we agreed on by the total number of items and then multiplying by 100. Based on previous research, students with LD tend to score a 0 or a 1 on the final examiner judgment. When dichotomizing the final examiner judgment where 0 or 1 = LD and 2, 3, or 4 = typical development (TD), inter-rater reliability was 92%. Table 3 shoes the percent agreement for each individual measure.

Table 3

Inter-Rater Reliability for DA Pretest, Posttest, and Modifiability Measures

Static measure	% agreement	Modifiability measure	% agreement
Pretest 1 & Teaching Phase 1	96%	Response to Prompts	68%
Posttest 1	92%	Degree of Transfer	68%
Teaching Phase 2	88%	Attention	64%
Posttest 2	88%	Ease	60%
Posttest 3	72%	Frustration	84%
		Disruptions	84%
		Mod Total	56%
		Final Judge	68%
		Dichotomized Final Judge	92%

Note. Mod Total = DA modifiability total score. Final Judge = final examiner judgement.

Discussion

The purpose of this study was to determine the sensitivity, specificity, and inter-rater reliability of a DA of inferential word learning. While our DA primarily used a test-teach-retest approach, it also incorporated an aspect of graduated prompting in that it provided a standardized teaching script with more instruction offered to students who needed more support. Traditionally, test-teach-retest models are less scripted and do not include a graduated prompting approach. However, we attempted to increase uniformity in test administration and incorporate a graduated prompting approach, which has been successful in previous vocabulary DA studies (Kapantzoglou et al., 2011). Consequently, we did not uniformly administer the DA between participants because the second teaching phase was optional and was administered based on the participants' responses.

We hypothesized that this DA would have superior sensitivity and specificity compared to that of static vocabulary measures and similar sensitivity and specificity compared to that of

Petersen et al. (2019). Results of this study indicate that the sensitivity and specificity of the DA of inferential word learning were acceptable, but lower than that found in Petersen et al. (2019). We also hypothesized that the DA of inferential word learning would have high interrater reliability. We found that the static measures (with the exception of posttest three) and the dichotomized final examiner judgement had excellent reliability, while the individual modifiability measures (with the exception of disruption and frustration) had poor reliability.

Sensitivity and Specificity

We found that posttests 1, 2, and 3 did a poor to fair job at predicting LD. When we added the total modifiability score to the posttests, the specificity and sensitivity increased from fair to good. Interestingly, when we added the final examiner judgement to all of the posttests and the modifiability total (Model 1, Step 2) the sensitivity and specificity remained relatively unchanged. Previous studies have found that combining modifiability ratings and posttest scores generally yield the highest sensitivity and specificity values (e.g., Kapantzoglou et al., 2011; Peña & Lidz, 2001; Peña, Reséndiz, & Gillam, 2007; Petersen, Chanthongthip, Ukrainetz, Spencer, & Steve, 2017; Ukrainetz et al., 2000). Likewise, combining all posttests, the modifiability total, and the final examiner judgement scores from this DA yielded the strongest sensitivity (.83) and specificity (.80). The classification accuracy of this DA was lower than that of Petersen et al. (2019; sensitivity = 90% and specificity = 90.5%), but higher than many static norm-referenced vocabulary tests (Gray et al., 1999). Petersen et al. (2019) had a much smaller sample size (N = 31) compared to our study (N = 127). Because our study had a larger sample size, our sensitivity and specificity may be more accurate. However, Petersen et al.'s (2019) DA included a teaching phase that targeted inferential word learning in sentences. In their study, some students with LD were able to define words after listening to stories. These students'

language difficulties only manifested in the context of defining words in sentences. Our DA only taught students to define words in story contexts. Thus, our test would not have identified students with LD who are able to infer word meaning from stories but not sentences. Additionally, Petersen et al. (2019) administered the DA to students in kindergarten through third grade. Our study included students from kindergarten through fifth grade. It could be that a child's ability to learn word meaning using context clues is a more appropriate diagnostic marker for children in earlier grades. Nevertheless, the classification accuracy of the DA used in this study suggests that it is worth developing and researching further. This study adds to the current body of research by reaffirming that the ability to learn vocabulary using context clues is a strong indicator of LD.

Inter-Rater Reliability

We expected the DA pretest and posttest static measures to have high reliability because of their objective nature and clear scoring procedures. On three of the static measures (pretest and posttests 1 and 2), we had excellent reliability (88-96%), but posttest 3 had lower reliability than expected (72%). Posttest 3 was the sentence-based test item. In the future, scoring instructions for the sentence-based posttest should provide more examples of correct and incorrect responses to potentially increase agreement between examiners.

We expected the modifiability measures to have lower reliability compared to the pretest and posttest measures because they are subjective ratings. With the exception of disruption (84%) and frustration (84%), our individual modifiability measures yielded poor reliability (56-68%). Previous research asserts (Petersen et al., 2017) that students with LD tend to score a 0 or a 1 on the final examiner judgment. For this reason, we dichotomized the final examiner judgement—a score of 0 or 1 = LD and a score between 2 and 4 = TD. After dichotomization,

the interrater reliability was 92%. We consider the inter-rater reliability of the dichotomized final examiner judgement to be most clinically useful because the purpose of this DA is to identify children with and without LDs.

One objective of DA is to provide therapeutically relevant, qualitative information about the child's learning. Considering this objective, we think it is important to postulate why our inter-rater agreement was low on some of the individual modifiability items. After the administration of the DA to participants, our research team expressed concern that the modifiability scale poorly described learning behaviors that the participants exhibited during the DA; our reliability data supports their concerns. Our modifiability scale differed significantly from Petersen et al. (2019). In that study, modifiability scale items 1-6 consisted of statements (e.g., "the child attended to the teaching/testing") and the examiner rated the frequency of the behavior (2 = most of the time, 1 = some of the time, and 0 = none of the time). On the rating scale for the DA in this study, the scoring criteria for each item changed. For example, a score of 4 for "attention to teaching" is described as "attentive and focused. No verbal reinforcement needed." Whereas a score of 4 for disruptions is described as "little behavior that interrupts intervention." Additionally, for the final examiner judgement on the Petersen et al. (2019) DA, examiners responded to "What is your overall judgment of the child's potential to learn vocabulary through listening to stories?" (2=good, 1=average, 0=poor). However, the final examiner judgement in the DA used in this study, did not provide any instructions or descriptors. The modifiability scale used in this study was borrowed from the Petersen et al. (2017) dynamic assessment of narratives, and there appeared to be an improper application to the DA of inferential word learning used in this study. Decreasing ambiguity in our modifiability index will likely increase agreement between examiners.

Limitations and Future Research

As part of our reference index to establish LD, we administered a dynamic assessment of narrative language first, then administered the DA of inferential word learning. On the DA of narrative language, students were asked to listen to a narrative and then were taught how to retell that story. On our test, when given a prompt to define a word, some students retold the story instead of defining the new word. This may indicate that some children had difficulty transitioning between the assessments rather than having difficulty learning words through the use of context clues. In a future study we should administer this assessment first, or on its own to avoid this confounding variable.

Our ratio of participants with and without LD is not representative of the actual population. Data from a large epidemiological study indicated that the prevalence of LD is 7.4% (Tomblin et al., 1997). In our sample, 24% of the participants had an LD. According to Orellana et al. (2019), overrepresenting the ratio higher than the actual prevalence can inflate the classification accuracy. It is possible that our classification accuracy is inflated because children with LDs were overrepresented in our cohort.

An additional limitation of our study is the possibility of incorporation or verification bias. We decided how to classify our participants into LD or TD groups after administering the whole reference index battery and so it is possible that our index measure influenced our initial classification of LD. However, researchers were blind to student performance on the DA when establishing the reference index criteria.

While we reported steps we took to insure fidelity, we did not calculate percent adherence to the DA protocol or report any deviations from the protocol. Also, it might have been helpful to have each DA session video recorded so that both fidelity and inter-rater

reliability could be evaluated in greater detail. Future research should carefully document fidelity of the DA administration and use video recordings or real-time observation to examine inter-rater reliability.

While our study had its limitations, overall the DA of inferential word learning has superior sensitivity and specificity compared to static norm referenced vocabulary tests. It also adds to the current literature that suggests that inferential word learning is a valuable diagnostic marker for language disorders and that DA is an appropriate tool to measure children's ability to learn word meaning through the use of context clues.

References

Baumann, J. F., & Kameenui, E. J. (1991). Research on vocabulary instruction: Ode to Voltaire. In J. Flood, J. J. D. Lapp, & J. R. Squire (Eds.), *Handbook of research on teaching the English language arts* (pp. 604–632). New York, NY: MacMillan.

Camilleri, B., & Botting, N. (2013). Beyond static assessment of children's receptive vocabulary: A dynamic assessment of word learning ability. *International Journal of Language and Communication Disorders, 48*, 565-581.

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *International Journal of Speech-Language Pathology, 9,* 312-322.

Camilleri, B., & Law, J. B. (2014). Dynamic assessment of word learning skills of pre-school children with primary language impairment. *International Journal of Speech-Language Pathology, 16*, 507-516.

Dunn, L., Dunn, L., Whetton, C., & Burley, J. (1997). *The British Picture Vocabulary Scale*. Windsor, England: NFER-Nelson.

Dunn, Lloyd M., & Dunn, Leota, M. (1997). *Peabody Picture Vocabulary Test (3rd ed.)*. Circle Pines, MI: American Guidance System.

Feuerstein, R., Rand, Y., Hoffman, M. B. (1979). *The dynamic assessment of retarded Performers: the learning potential assessment device theory, instruments, and Techniques.* Glenview, IL: Scott, Foresman.

Gardner, M. F. (1985). *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications.

Gardner, M. (1990). *Expressive One-Word Picture Vocabulary Test, Revised.* Novato, California: Academic Therapy Publications.

Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*, 887–898.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196-206.

Gutiérrez-Clellen, V. F., & Peña, E. (2001). Dynamic assessment of diverse children: A tutorial. *Language, Speech, and Hearing Services in Schools, 32,* 212-224.

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

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. (2011). Dynamic assessment of word learning skills: Identifying language disorder in bilingual children. *Language, Speech, and Hearing Services in Schools, 43,* 81-96.

Larsen, J., & Nippold, M. (2007). Morphological analysis in school-age children: Dynamic assessment of a word learning strategy. *Language, Speech, and Hearing Services in School, 38*, 201-212.

Martin, N., & Brownell, R. (2012a). *Expressive One-Word Picture Vocabulary Test–Fourth Edition: Spanish (EOWPVT-IV: Spanish)*. Novato, California: Academic Therapy.

Publications Martin, N., & Brownell, R. (2012b). *Receptive One-Word Picture Vocabulary Test–Fourth Edition: Spanish (ROWPVT-IV: Spanish).* Novato, California: Academic Therapy Publications.

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

Nation, P., & Waring, R. (1997). *Vocabulary size, text coverage and word lists.* In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition, and pedagogy* (pp. 6-19). Cambridge, United Kingdom: Cambridge University Press.

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.

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.

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

Peña, E. D., Reséndiz, M., & Gillam, R. B. (2007). The role of clinical judgments of modifiability in the diagnosis of language disorder. *Advances in Speech-Language Pathology, 9*, 332–345.

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 Hearing Research, 60*, 983-998.

Petersen, D. B., & Spencer, T. D. (2016). *CUBED*. Language Dynamics Group.

Petersen, D. B., Tonn, P., Spencer, T. D., & Foster, M. E. (2019). The classification accuracy of a dynamic assessment of inferential word learning for bilingual English/Spanish-speaking school-age children. *Language, Speech, and Hearing Services in Schools, 51,* 1-21.

Ram, G., Marinellie, S. A., Benigno, J., & McCarthy, J. (2013). Morphological analysis in context versus isolation: Use of a dynamic task with school-age children. *Language, Speech, and Hearing Services in Schools, 44,* 32–47.

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.

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, Language, and Hearing Research, 40*, 1245–1260.

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

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Wolter, J. A., & Pike, K. (2015). Dynamic assessment of morphological awareness and third-grade literacy success. *Language, Speech, and Hearing Services in Schools, 46*, 112–126.

APPENDIX A

Dynamic Assessment of Inferential Word Learning (DA-IWL) Scoring Sheet

Phase	**Points**		
A. Pretest, Procedure 1 ***Benel***	2	1	0
teaching phase 1	2	1	0
B. Posttest 1 ***Tubik***	2	1	0
Posttest 1 teaching phase	2	1	0

Modifiability Scale. The examiner is to complete this modifiability rating scale immediately after the administration of the dynamic assessment. A Total Modifiability Index (TMI) is calculated by adding the points awarded for each of the seven areas.

MODIFIABILITY

POINTS	4	3	2	1	0
Response to prompts	Student responds to prompts **most** of the time (mostly level 1)		Student responds to prompts **some** of the time (mostly level 2)		Student responds to prompts **infrequently** (almost all level 2)
POINTS	4	3	2	1	0
Degree of transfer	**1-2 targets** are transferred across steps. All story grammar elements are included in final step.		At least **1 target** is occasionally transferred across steps. Many story grammar elements are included in the final step.		Transfer of targets is **rare** across steps. Few story grammar elements are included in the final step.
POINTS	4	3	2	1	0
Attention to teaching	**Attentive** and **focused**. No verbal redirections needed.		On task **some** of the time. Some verbal redirections needed and can be refocused.		**Distracted** and **difficult** to refocus. Significant redirection needed.
POINTS	4	3	2	1	0
Ease of teaching	**Minimal effort** required from the examiner to induce change. Examiner effort greatly decreases across steps.		**Some effort** from the examiner required to induce change. Examiner effort decreases somewhat across steps.		**Considerable effort** from the examiner required to induce change. Effort decreases very little across steps.
POINTS	4	3	2	1	0
Frustration	**Little** to **no frustration** exhibited. Persistent and engages in tasks readily.		**Some frustration** indicated. Tentative and unsure. May require soothing.		**Considerable frustration** exhibited. Distressed and requires significant soothing.
POINTS	4	3	2	1	0
Disruptions	**Little behavior** that interrupts intervention.		**Some behavior** that interrupts intervention.		**Considerable behavior** that interrupts intervention.

Circle the number that reflects the student's overall responsiveness during the teaching phase.

MODIFIABILITY SCORE = 24

FINAL EXAMINER JUDGEMENT

POINTS	4	3	2	1	0
Final Examiner Judgement of Student's Ability to Learn Vocabulary Through Listening to Stories					

FINAL EXAMINER JUDGEMENT SCORE= 4

Posttest 2 ***Tanif***	2	1	0
Posttest 3, sentence. ***Glistern***	2	1	0
Total Score (not including modifiability scores)			

APPENDIX B

Annotated Bibliography

Camilleri, B., & Botting, N. (2013). Beyond static assessment of children's receptive vocabulary: A dynamic assessment of word learning ability. *International Journal of Language and Communication Disorders, 48*, 565-581.

Objective: This study investigated the reliability and correlational validity of Dynamic Assessment of Word Learning (DAWL).

Method: Participants consisted of 15 nursery students ages 3-4 who were referred for speech-language services. Researchers measured non-verbal cognitive skills using the Building Block task from British Abilities Scale II (BAS) (Elliott, 1996)]. Researchers then administered the British Picture Vocabulary Scale II (BPVS-II) to gauge the participant's receptive vocabulary. Next, the participant and researcher viewed and talked about pictures together. Two of the pictures contained target words which the individual participant missed on the BPVSII. The examiner used a prompting hierarchy to cue the child's identification of the target word in the second presentation and the ability of the child to use the target expressively was scored. The posttest was similar to the interactive phase, but the target words were words considered above preschool ability.

Results: This administration of the Dynamic Assessment of Word Learning (DAWL) resulted in appropriate internal consistency, inter-rater & test-retest reliability as well as correlational, concurrent, and predictive validity

Relevance to current work: Results suggest that the DAWL can complement static tests.

Camilleri, B., & Law, J. (2007). Assessing children referred to speech and language therapy: Static and dynamic assessment of receptive vocabulary. *International Journal of Speech-Language Pathology, 9,* 312-322.

Objective: This pilot study compared the ability of Dynamic Assessment (DA) of word learning to classify children with typical and atypical language skills to that of a static vocabulary measure. It also investigated the difference between monolingual English speakers and multilingual students' scores on both static and dynamic measures.

Method: Camilleri et al. developed a DA from the BPVS-II (Elliott, 1996) and administered the assessment to 54 preschoolers. Researchers suspected 40 of the children had language deficits. Fourteen children had typical language abilities and 12 of the total participants spoke an additional language. The students took the British Picture Vocabulary Scale II and six deficit target words were targeted in the during a play-based teaching phase.

Results: Children who spoke an additional language scored lower on the BPVSII than their monolingual English speaker peers. However, this difference did not exist in the BAS score or the dynamic measures. The children with suspected typical language development scored significantly higher on all measures they participated in.

Relevance to current work*:* This study contrasts the lingual bias of DA to that of static vocabulary assessments.

Camilleri, B., & Law, J. B. (2014). Dynamic assessment of word learning skills of pre-school children with primary language impairment. *International Journal of Speech-Language Pathology, 16*, 507-516.

Objective: This the purpose of the article is to determine the degree to which DA and SA scores predict vocabulary performance over time and if there was a different pattern of prediction for low scoring children vs. children referred for evaluation as a whole.

Method: 50 inner city children aged three to five were referred for speech language services because of concerns of speech and language. Children were divided into high scoring (BPVS-II >25th percentile) and low scoring groups (BPVS-II <25th percentile). Children retook the BPVS-II after six months. The testing procedures followed the procedures outline in Camilleri & Law's 2007 study. Six months after the initial testing, the children took the BPVS-II again

Results: 24 children were placed in a higher percentile on the BPVS-II at the second administration and 8 children were placed in a lower percentile. As a whole, the BPVS-II correlated with the dynamic measure, however, only one dynamic measure correlated with the second BPVS-II score for one participant in the low language ability group. For low scoring children, the static BPVS-II had low predictive validity on the second BPVS-II administration.

Relevance to current work: This study suggests that DA is more predictively relevant for children with low language abilities.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196-206.

Objective: This study was designed to evaluate the validity of common vocabulary tests, their ability to separate SLI children from NL children, and the validity of interpretation of these scores.

Methods: Thirty-one pre-school age children (4- and 5-year-olds) had SLI and 31 had typical language. Researchers administered the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), Receptive One-Word Vocabulary Test (Gardner, 1985) and Expressive One-Word Vocabulary Test-Revised (Gardner, 1990) to each child.

Results: Results indicated that the students with SLI performed significantly lower than the students with typical language, with strong effect sizes. Also, the vocabulary tests all significantly correlated with each other. The participants with typical language regularly scored higher than the children with SLI on each test. However, the majority of children with SLI scored within one standard deviation of the mean, indicating normal language. These vocabulary tests have construct validity, but they are not valid indicators of SLI or NL. None of the tests stood out for identifying SLI. Using multiple vocabulary tests will not allow for accurate diagnosis of SLI.

Relevance to current work: These vocabulary tests have construct validity, but they are not valid indicators of SLI or NL.

Gutiérrez-Clellen, V. F., & Peña, E. (2001). Dynamic assessment of diverse children: A tutorial. *Language, Speech, and Hearing Services in Schools, 32,* 212-224.

Objective: This the purpose of the article is to compare DA methods in their ability to separate language difference and disorder and to present a DA protocol that minimizes misdiagnosis. Additionally, the article presents a case study that shows the ability of DA to separate disorder from difference compared to traditional norm referenced tests.

Method: Researchers chose two Latin American, bilingual, Spanish-English speakers, students in from the same Head Start program class. Researchers determined the participants' language abilities through classroom observation, parental & teacher

reports, and standardized assessments—Expressive One Word Picture Vocabulary Test-Revised (EOWPVT-R) Comprehension subtest of the Stanford-Binet Intelligence Scale, and 10 items (5 expressive and 5 receptive) of the Preschool Language Scale. They were taught strategies to use single word labels in 2 30-miute sessions.

Results: Researchers found that CLD children made more progress regarding pretest and posttest scores than their non-CLD peers. Consequently, looking at the pretest and posttest scores individually provides more valuable information than a gain score.

Relevance to current work: Teach-test-retest is an effective way to differentiate difference and disorder in language disorders. It is effective because the examiner does not come in with bias from prior knowledge of the student's ability. Hierarchy of prompts does not serve as a diagnostic indicator because it does not directly teach the student the problem-solving skills. The author stipulates that the posttest score and the modifiability rating provide the most valuable information.

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. (2011). Dynamic assessment of word learning skills: Identifying language disorder in bilingual children. *Language, Speech, and Hearing Services in Schools, 43,* 81-96.

Objective: The researchers wanted to know if applying DA to word learning skills would accurately identify primary language disorder in children who spoke both English and Spanish.

Method*:* The participants consisted of 28 preschoolers with Spanish as their primary language (15 TD, 13 LI). Children with better expressive skills in Spanish than English were presented with 3 unfamiliar objects & 3 familiar objects. The researchers gave each object a nonword CVCV label that matched that consisted of early developing

consonants in the Spanish Language. The teaching phase consisted of an individualized "scripted structured play activity" with nine presentations of the target words and three opportunities for the participant to produce the target word correctly.

Results: Children with typically developing language learned words faster than children with Primary Language Disorder. A total of 78.6% were classified correctly with 76.9% sensitivity and 80% specificity.

Relevance to current work: DA is likely a valid way of assessing word learning skills.

Larsen, J., & Nippold, M. (2007). Morphological analysis in school-age children: Dynamic assessment of a word learning strategy. *Language, Speech, and Hearing Services in School, 38*, 201-212.

Objective: The purpose of this study was to show the efficacy of The Dynamic Assessment Task of Morphological Analysis (DATMA).

Method: Fifty typically developing 6th grade monolingual English speakers took the Dynamic Assessment Task of Morphological Awareness (DATMA), Oregon Statewide Assessment (OSA), and Peabody Picture Vocabulary Test (PPVT.) The DATMA consists of 15 Low frequency derivations of high frequency words. The examiner used graduated prompting to determine the level of adult support a child needs in order to complete the task. Each child was given an overall literacy score which combined the child's OSA and PPVT scores. Then the researchers compared the participant's overall literacy and DATMA scores.

Results: Test resulted in distinct literacy levels. Student's scores on the DATMA correlated with their OSA and PPVT scores. Most children required 1 or 2 prompts to get an acceptable score.

Relevance to current work: The researchers applied dynamic assessment to vocabulary assessment. Modifiability score was the strongest separating factors (confirms other studies).

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.

Objective: This study looked at dynamic assessment with a word learning task. Main research questions included 1. Does DA have a lower rate of false identification of CLD children as language impaired compared to static normed tests? 2. Do children with different language abilities respond differently to short-term MLEs? 3. Do effects of MLE transference to other tasks?

Method: Sample population: 79 Bilingual head start children with a mean age of 4;2 with a varied use of English and/or Spanish. Children were divided into two groups—Low Language ability and Typical development. Tests administered: a modified version of the PLS, EOWPVT, and Comprehension subtest of the Stanford-Binet Intelligence Scale. Test instructions were given in the child's home language. Responses in either language were accepted. MLE consisted of two 30-minute sessions spaced 1-2 weeks apart with the goal of teaching single-word labeling. Instruction also included planning and self-regulation strategies. The control group did not participate in a teaching phase.

Results: (a) CLD participants with TD language scored improved their scores. The second administration of the tests had higher overall classification rates. CCSB was the exception (it separated the children's language abilities in the pre-test) (b) Children in the TD group increased their post-test score while LLA kids scores remained the same. (c) After mediation, TD participants increased their scores on all language assessments. Out of all the test combinations, the post-test EOWPVT score with the modifiability rating had the highest classification accuracy (95.3% correct TD classification) and specificity rate of 77.8% (LLA to LLA group)

Relevance to current work: It taught the CLD children academic skills necessary to perform well on standardized tests.

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

Objective: The purpose of this study was to show the efficacy of DA in the assessment of children with Language Disorders.

Method: 50 students in Spanish-English bilingual head-start classrooms, aged 3-9 years, participated in this study. Students were placed in a possibly language disordered (PLD) or a non-disabled group based on observations of their interactions with peers or teacher or parent report, the EOWPVT, Comprehension subtest of the Stanford-Binet Intelligence Scale. In the pretest, the majority of students (44/50) scored lower than 90 on the EOWPVT, indicating a risk for language learning difficulty. The mediation phase comprised of two 20-minute small group interventions that focused on labeling strategies.

Results: The PLD group scored significantly lower on the pretest CSSB than the non-disabled children, but both groups scored similarly on the EOWPVT. With the

combination of the EOWPVT2 and MI, 93% of the children were correctly identified. Both groups scored higher on the second administration of the EOWPVT. However, the non-disabled group made significantly larger gains in their scores.

Relevance to current work: This study is important because it shows that while one administration of a static vocabulary test misidentifies children with diverse linguistic backgrounds and that test-teach-retest. When you give the test a second time and compare the modifiability index and the second test score, you get much higher accuracy.

Petersen, D. B., Tonn, P., Spencer, T. D., & Foster, M. E. (2019). The classification accuracy of a dynamic assessment of inferential word learning for bilingual English/Spanish-speaking school-age children. *Language, Speech, and Hearing Services in Schools, 51,* 1-21.

Objective: The purpose of this study was to compare the ability of a dynamic assessment of inferential word learning to differentiate between language disorder and language difference in bilingual children to that of static vocabulary assessment.

Methods: Thirty-one Spanish/English bilingual school aged students participated in this study. Researchers administered the Expressive/Receptive Expressive One Word Vocabulary Test, Frog Retell, and the dynamic assessment.

Results: The dynamic assessment had higher sensitivity (100%) and specificity (95%) when combining the posttest and modifiability ratings compared to that of the static assessments.

Relevance to current work: This is a preliminary report of DA-IWL which reports high classification accuracy.

Ram, G., Marinellie, S. A., Benigno, J., & McCarthy, J. (2013). Morphological analysis in context versus isolation: Use of a dynamic task with school-age children. *Language, Speech, and Hearing Services in Schools, 44,* 32–47.

Objective: This study aimed to determine if the ability of third and sixth graders to use a morphological strategy to determine word meaning differed, to investigate the effect of presenting the word in isolation vs context, and to determine if the participant's DA score will correlate with reading volume and print exposure.

Method: children in grades three and five defined words for researchers from a modified version of the DATMA. The researchers administered 20 Low frequency derivations of high frequency words (nouns or adjectives) both in isolation and in context. Test administers used a prompting hierarchy to determine the amount of adult support necessary for the child to determine word meaning.

Results: Children needed less prompts to determine word meaning when the target word was presented in context vs in isolation. Additionally, younger children used context clues over morphological analysis to determine word meaning. The children's ability to determine word meaning strongly correlated with their reading frequency.

Relevance to current work: One weakness of this study is that morphological analysis requires previous knowledge of root words, suffixes, and their relationship to grammatical classes. Additionally, some participants may have had familiarity with some of the words included such as "guitarist, hurtful, and dressy," Using non-words may help mitigate these factors.

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.

Objective: Spaulding, Plante, and Farinella conducted a systematic review of vocabulary tests. They investigated whether the Peabody Picture Vocabulary Test-III (Dunn & Dunn, 1997), Receptive One-Word Vocabulary Test (Gardner, 1985), Expressive Vocabulary Test (Williams, 1997), and Expressive One-Word Vocabulary Test-Revised (Gardner, 1990) all measure the same construct for children with and without language impairment (convergent evidence of validity), whether there was evidence of evidence of divergent validity, and sensitivity and specificity.

Method: The researchers reviewed the latest editions of 43 norm-referenced standardized tests that claim to "test English language skills" to "identify childhood language impairments". They excluded tests that looked at academic skills, interviews, or observing the child and most criterion-referenced tests and screeners. They collected data published in the test manuals that was explicitly provided or provided information that allowed the researchers to easily calculate sensitivity and specificity.

Results: Results indicated that the students with SLI performed significantly lower than the students with typical language, with strong effect sizes. Also, the vocabulary tests all significantly correlated with each other. There was evidence of divergent validity based on low correlations with factors no expected to influence vocabulary performance. However, sensitivity and specificity were low. Sensitivity ranging from 71% to 77% and specificity ranging from 68% to 77%. Although the participants with typical language regularly scored higher than the children with SLI on

each test, the majority of children with SLI scored within one standard deviation of the mean.

Relevance to current work: Vocabulary assessments in particular have poor classification accuracy.

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

Objective: This study investigated the ability of the post-test and modifiability score on a DA of categorization to differentiate Native-American children with strong and weak language abilities.

Method: Ukrainetz et al. used the receptive and expressive categorization subtests from Assessing Semantic Skills through Everyday Themes (ASSET; Barret, Zachman, & Huisingh, 1988) to measure categorization skills in a test-teach-retest model of dynamic assessment. Twenty-three Arapahoe and Shoshone English-speaking kindergarteners participated in this study. The researchers divided the children into two groups—strong language learners (n = 15) and weak language learners (n = 8) based on classroom observations and teacher reports. Each student received two 30-minute mediation sessions with a peer. The teaching phase targeted the overarching skill of grouping, rather than targeting specific vocabulary words. After each session, the clinician filled out the learning abilities scale and the response to mediation Likert scales for each child and combined these ratings into an overall modifiability index.

Results: The modifiability index significantly differentiated between the strong language learners and weak language learners with a large effect size and a low

probability that the groups' scores overlapped. The response to mediation checklist (measure of transfer and responsivity) specifically differentiated strong language learners and weak language learners with higher accuracy than the learning abilities scale. Children rated as high language learning ability made greater pre/posttest gains compared to their low language learning ability peers.

Relevance to current work: This study supports the use of DA as a more effective measure of language ability in children from culturally and linguistically diverse backgrounds than static norm-referenced tests. More specifically, this study supports the efficacy of measures of transfer and responsivity to differentiate between children of high and low language abilities.

Wolter, J. A., & Pike, K. (2015). Dynamic assessment of morphological awareness and third-grade literacy success. *Language, Speech, and Hearing Services in Schools, 46*, 112–126.

Objective: This study built on the findings of Larsen and Nippold's 2007 and Ram et Al.'s 2013 studies in a lower grade population.

Method: Fifty-four typically developing, monolingual English speaking, homogenous third graders participated in this study. To measure morphological awareness, the researchers modified Nippold's DATMA to an age-appropriate level. The researchers also administered assessments to measure a variety of literacy skills—phonemic awareness (CTOPP), receptive vocabulary (PPVT-4), sight-word reading (WRMT-R WID), decoding (WRMT-R WA), reading comprehension (WRMT-R PC), and spelling (TWS-4).

Results: Children's DAPMA scores significantly correlated with their reading comprehension skills. Third graders' DAPMA scores resulted in a standard bell curve

indicating normality (normal score distribution, small kurtosis and skewness). The large majority of students were unable to define the target words without adult scaffolding. Additionally, Wolter et al. reported lower interrater reliability compared to that of Ram et al.'s study.

Relevance to current work: This study suggests that the dynamic assessment is a valuable tool for measuring early derivational morphological awareness skills in third graders. This study also highlights that in lower grades, derivational morphological awareness skills are emerging as evidenced by the high amount of adult scaffolding required for the students to identify word meaning. Thus, a measure of children's ability to infer word meaning using context may be more appropriate for measuring vocabulary skills in children in young grades because this skill typically develops earlier than morphological awareness. Additionally, the researchers hypothesized that a measure of contextualized word-learning may result in higher interrater reliability in younger grades because coding the correctness of student's responses is easier and more straightforward.