2021-04-08

Norming a Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Preliminary Psychometric Study

Ashley Elizabeth Frahm
Brigham Young University

Norming a Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Preliminary Psychometric Study

Ashley Elizabeth Frahm

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
Garrett Cardon

Department of Communication Disorders

Brigham Young University

ABSTRACT

Norming a Dynamic Assessment of Narrative Language for Diverse
School-Age Children With and Without Language Disorder:
A Preliminary Psychometric Study

Ashley Elizabeth Frahm
Department of Communication Disorders, BYU
Master of Science

The purpose of this study was to investigate preliminary psychometric normative data of an English dynamic assessment of narrative language for a group of diverse school-age students with and without language disorder. This study included 364 diverse students with and without language disorder ranging from kindergarten through 6th grade. Students were confirmed as having a language disorder if they had an existing active IEP for language, and scores below a certain cutoff point on a nonword repetition (NWR) task and the narrative language measure (NLM). English language proficiency was investigated, and students were classified as being a dual language learner (DLL) based on student, teacher, or parent report of diverse home language, and poor performance on an English narrative language assessment. Participants were administered a nonword repetition task (NWR), the Narrative Language Measure (NLM), and the Dynamic Assessment of Oral Narrative Discourse (the DYMOND). Data were analyzed within groups of typically developing students and students with a language disorder to identify statistically different mean modifiability and posttest scores given various demographic factors. Results of this study indicate that modifiability and posttest scores for typically developing students were not found to be statistically different given gender or school location, however, significant differences were noted given grade and level of English proficiency or DLL status. The group of students with a disorder demonstrated no statistically different mean modifiability scores given any demographic factor. Students with a language disorder demonstrated significantly different mean posttest scores given school location and English proficiency and DLL status. Results from this study are consistent with previous dynamic assessment research in demonstrating excellent classification accuracy in culturally and linguistically diverse (CLD) populations. Students may benefit from a norm-referenced dynamic assessment of narrative language in order to provide less-biased standardized forms of assessment for CLD populations.

Keywords: dynamic assessment, narrative language, norm-referenced, language disorder, culturally and linguistically diverse (CLD) populations

ACKNOWLEDGMENTS

The author would like to thank the following individuals for their invaluable contribution as mentors and committee members: Douglas B. Petersen, Garret Cardon, and Connie Summers. Their knowledge, experience, and mentorship were crucial to the success of this project. The author also thanks the many members of the DYMOND research team for their dedication and effort in the data collection and analysis required for this project, and her husband, parents, family, and friends for their continual support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *Norming a Dynamic Assessment of Narrative Language for Diverse School-Age Children With and Without Language Disorder: A Preliminary Psychometric Study*, is written in journal publication style format. The preliminary pages of the thesis are consistent with university submission requirements. The body of the thesis is presented as a journal article and conforms to length and style requirements for submission to education journals, following APA formatting requirements. Portions of this manuscript may be altered and submitted for publication in a peer-reviewed journal with the primary author listed as a contributing author. Appendix A consists of an annotated bibliography. Appendix B consists of an approval letter from the IRB.

Introduction

The United States Census Bureau estimates that 21.9% of households in the US speak a language other than English in the home (United States Census Bureau, 2018). This cultural and linguistic diversity in US homes makes assessment and identification for language disorder in children in the US school system a complex task. Service providers are hard tasked to keep up with the rapidly increasing diversity in American schools. Of the 201,961 professionals representing the American-Speech-Language-Hearing Association (ASHA), only 6.5% met criteria for a bilingual service provider, with the majority indicating Spanish-English bilingualism (ASHA, 2019). The Individuals with Disabilities Education Act (IDEA, 2004) specifies the right for every child, regardless of disability or English proficiency to receive "early identification and assessment of disabling conditions" (p. 856). For speech-language pathologists (SLPs), this means every child, regardless of gender, grade, age, disability, school location, English language proficiency or dual language learning (DLL) status has the right to timely and appropriate assessment of language, speech, and other areas within the scope of SLP practice. IDEA further recognizes the need for appropriate assessment for culturally and linguistically diverse (CLD) populations, supporting the "development and provision of appropriate alternate assessments that are valid and reliable for assessing the performance of children with disabilities" (IDEA, 2004, p. 864). It is the responsibility of each professional to determine the most appropriate, valid forms of assessment for each student in order to accurately diagnose language disorder.

Although alternate assessment approaches may be necessary to validly determine need for special education services, Betz et al. (2013) found that SLPs in elementary school settings preferred to use language assessments that have a strong reputation in the field and are recommended by fellow professionals, regardless of empirical evidence of validity and reliability. Often these popular, traditional tests are standardized, norm-referenced assessments. A survey of 130 SLPs providing language assessment services for bilingual children in Michigan indicated that a majority used a combination of norm-referenced tests (NRTs) and informal measures to assess bilingual children. Norm-referenced assessments were more frequently used, with a vast majority of tests administered in English (Caesar & Kohler, 2007). In fact, NRTs make up the majority of diagnostic and service-qualifying assessments in circulation for language disorder (Owens & Pavelko, 2017).

The popularity of standardized NRTs has arisen in the field of speech-language pathology because of their perceived objective nature, and their ability to situate a student in relative standing to their peers (McCauley & Swisher, 1984). For the majority of SLPs, NRTs are clearly perceived as a highly important diagnostic tool in assessing a student's language and in qualifying a student for language services in schools. A common rationale for the use of NRTs in school-based clinical practice is to comply with the Committee of Special Education (CSE) and the district or state policy or preferences regarding the qualification for special education services (Fulcher-Rood et al., 2018; Spaulding et al., 2006). The state of Missouri for example requires student scores on a minimum of 2 NRTs to fall below a certain standard deviation in order to qualify for school-based services (Selin et al., 2019). The state of Florida requires a minimum of one norm-referenced test to be included in the language assessment process to qualify a student for school-based SLP services. Furthermore, the student must have a "standard score significantly below the mean" to qualify for services (Exceptional Student Education, 2016). Given that NRT scores may be required or are strongly suggested by some school districts to qualify a student for services, SLPs should be careful to examine each assessment to ensure that it is an appropriate measure to use with the individual being assessed.

The American Psychological Association defines a Norm-Referenced Test (NRT) as "any assessment in which scores are interpreted by comparison with a norm, generally the average score obtained by members of a specified group" (APA, 2020). NRTs have several purposes including providing evidence of a problem, indicating the need for more in-depth assessment, and marking an individual's need for commencement or continuation of therapy services. What makes NRTs different from other forms of assessment is their ability to compare an individual's performance to a normative sample, typically made up of age-matched peers (McCauley & Swisher, 1984). SLPs often use NRTs in practice to diagnose language disorder by comparing a child's performance against that of a normative sample (Owens & Pavelko, 2017). Furthermore, the majority of NRTs currently in circulation are English static measures, meaning they test current performance or knowledge.

The very nature of English static measures makes them biased against students who are English language learners or those who may not be familiar with cultural aspects embedded in the assessment. Given that many NRTs are biased against CLD populations, they often misrepresent this population's actual abilities leading to overidentification of CLD children as having language impairment (Laing & Kamhi, 2003; Orellana et al., 2019; Owens & Pavelko, 2017). Because diversity is increasing across the United States, there is an urgent need to more accurately distinguish between those who have a language disorder and those who are culturally and linguistically diverse. An NRT's diagnostic accuracy, usually evidenced through sensitivity (i.e., percentage of students with language impairment identified as having language impairment) and specificity (i.e., percentage of students without language impairment identified as not having language impairment) should be established before it is used for the diagnostic decision-making process. In fact, Spaulding et al. (2006) recommended that first, diagnostic accuracy be confirmed, then other psychometric characteristics be examined. Among those psychometric characteristics, establishing evidence of construct validity and demonstrating that the normative sample represents the population for which the assessment is designed is crucial.

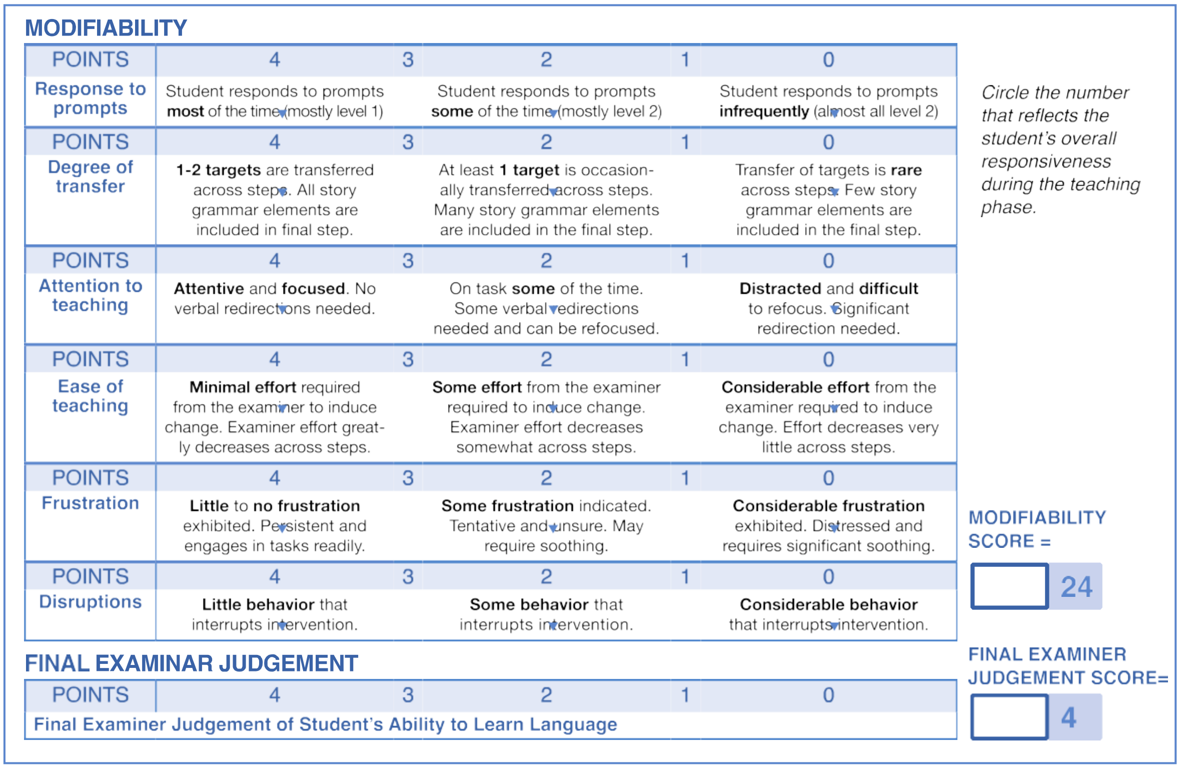
Hutchinson (1996) outlined the importance of determining a test's construct validity, or how well the test measures what it is expected to measure. Construct validity identifies separate test items and components to ascertain whether cumulatively, the assessment accurately reflects the test maker's intent or theory underlying the test. Construct validity can be evidenced in several ways. When results of a test show patterns of performance as the theoretical construct would indicate, such as when scores from a static NRT increase when student age increases, this is evidentiary of construct validity.

Peña, Spaulding, and Plante (2006) suggest that due to the nature of static NRTs, the normative sample should reflect the population that the test is designed to assess (see also Hutchinson, 1996). This is largely due to the fact that the norms reported in an NRT are used to interpret scores and student relative performance. Hutchinson (1996) maintained that reasonable representation of a normative sample should consider both adequate sample size and the type and demographics of participants. Language impairment is a disorder that appears to be normally distributed in the general population due to its polygenic nature (Tomblin et al., 1997) therefore, performance on a static NRT designed to identify language impairment in the general population should be normally distributed along the normal bell-curve when a sufficient representative sample of the population has been tested. Proportion is an important aspect to consider when establishing an accurately representative sample for an NRT. In order to allow for appropriate comparison of students in the general population to the normative sample, proportions of groups within the normative sample should match proportions of those groups that exist in the general population. If the test is meant to assess a more specific group within the population, its normative sample should include accurate and appropriate representation from that group (Hutchinson, 1996). For example, if the NRT is meant to assess the language of male and female students ages 5-12 with hearing loss, that assessment should be normed on an adequate number of male vs. female students with hearing loss, including a sufficient and matched number of students at every age from 5-12 years. These claims are logical for a test that is designed to measure the student's current level of performance.

Dynamic assessment (DA), as opposed to static assessment, is designed to measure a student's ability to learn, or in other words, how modifiable they are. It is derived from Vygotsky's (1978) research on the zone of proximal development, what the student is capable of when provided scaffolding just outside of their realm of ability. DA follows a test-teach-retest format, including pretest, modifiability, and posttest scores measuring current level of performance, the student's modifiability, and performance after receiving scaffolding within the student's zone of proximal development (Laing & Kamhi, 2003). Typically, measures such as ease of teaching, distractibility/disruption, responsiveness to teaching, degree of transfer of knowledge, and student frustration level are rated with a modifiability Likert scale (see Figure 1) that is used to help determine how well the student learned during the teaching phase (Petersen et al., 2020). Modifiability ratings in a DA are considered less biased against CLD populations as they should not be impacted by a student's gender, grade, school location or level of English proficiency or DLL status. Alternatively, posttest ratings in a DA may vary based on factors such as a student's grade, English proficiency or DLL status, simply given that students who are older or have higher English proficiency are likely to include more language complexity and story elements than their younger peers or those who have less experience with the English language. Many DA studies to date indicate that the best diagnostic accuracy for DA is achieved when both modifiability measures and posttest scores are considered (Peña, Gillam et al., 2006; Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020; Ukrainetz et al., 2000). This form of assessment can be a valuable language diagnostic tool in that all studies on DA to date have indicated good to excellent sensitivity and specificity (Kapantzoglou et al., 2012; Kramer et al., 2009; Peña 2000; Peña, Gillam et al., 2006; Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020; Roseberry & Connell, 1991; Ukrainetz et al., 2000).

Figure 1

DYMOND Modifiability Rating Scale



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 EXAMINAR JUDGEMENT

POINTS	4	3	2	1	0
Final Examiner Judgement of Student's Ability to Learn Language					

FINAL EXAMINER JUDGEMENT SCORE= □ 4

English DA of narrative language in particular is a valuable tool in the assessment of language for CLD populations (Orellana et al., 2019). Children from CLD populations are often at a disadvantage when participating in static testing due to lack of exposure to linguistic or cultural aspects that may be present in test items on the assessment, or the testing process in general. In contrast, an English DA of narrative language is considered to be a less biased form of assessment as its primary function is to assess the student's ability to learn a new aspect of language such as story grammar elements rather than testing what they currently know (Orellana et al., 2019). Along with culture and language differences, DA is a less-biased form of assessment when it comes to individual life experiences and academic knowledge (Laing & Kamhi, 2003).

Given that DA of narrative language (a) has strong evidence of diagnostic accuracy, (b) is a less-biased form of assessment for the increasing CLD population in the U.S., and (c) NRTs are required or strongly encouraged to qualify students for school-based special education services, it may be beneficial to develop an English norm-referenced DA of narrative language. Contrary to established conventions for static normative samples, a normative sample for a norm-referenced DA of narrative language may not need to have proportionate racial and ethnic representation. Specifically, the modifiability measures from a DA should not differ according to gender, grade, school location, and English language proficiency. This is because modifiability, which measures innate learning ability, is not determined by linguistic, social, or cultural demographics (Feuerstein, 1979). Furthermore, a student's ability to learn the basic narrative elements as targeted in a DA of narrative language at posttest may be less connected to demographic variables such as gender and school location (rural/urban). However, it is likely that a student's level of exposure to the English language and grade level factors would influence their ability to adequately recount a narrative within groups of children who are typically developing and those who have a language disorder.

The purpose of this study is to examine the preliminary psychometric normative data for an English dynamic assessment of narrative language for a diverse group of kindergarten through sixth grade students with and without language disorder. Specifically, the hypotheses are as follows:

1. It is hypothesized that mean modifiability scores will be equivalent for children who are typically developing regardless of grade (K-6), gender (male/female), school location (rural/urban area), and English language proficiency/exposure (proficient/dual language learner (DLL)).
2. It is hypothesized that mean modifiability scores will be equivalent for children who have language disorder regardless of grade, gender, school location, and English language proficiency/exposure.
3. It is hypothesized that within the group of typically developing students, there will not be any significant differences in posttest scores for gender, and school location, but that there will be a significant difference between students of varying grades, and between those who are English language proficient and students who are DLLs.
4. It is hypothesized that within the group of students with a language disorder, there will not be any significant differences in posttest scores across gender, and school location, but that there will be a significant difference between students of varying grades, and between those who are English language proficient and students who are DLLs.

Method

Participants

Participants in this study included 364 diverse Kindergarten through 6th grade students with and without language impairment, attending 4 elementary schools in Utah, Colorado, and Wyoming. Students were identified as having a language disorder when they met all three of the following criteria: (1) an active IEP for language services confirmed by the school SLP, (2) a nonword repetition (NWR) (CNRep; Gathercole et al., 1994) task score less than or equal to 71% accuracy, and (3) a narrative language measure (NLM) score of –1.5 standard deviations or lower (using national norms for grades K-3 and sample-specific data for grades 4-6). Demographic information regarding the sample such as gender, grade, school location, English language proficiency, DLL status, and disorder vs. no disorder are reported in Tables 1 and 2. Students were identified as being dual language learners (DLL) if the student, parent or teacher reported the student's home language as a language other than English and their NLM score in English was below the sixteenth percentile.

Table 1

Demographic Information for Total Sample

Demographic Group	*n* (%)
Total 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%)
English Proficiency	
Dual Language Learner (DLL)	31 (8.6%)
English Proficient	331 (91.4%)

n = Sample Size

Table 2

Demographic Information for Students With and Without Language Disorder

Demographic Group	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%)
English Proficiency		
Dual Language Learner (DLL)	22 (7.1%)	9 (16.7%)
English Proficient	286 (92.9%)	45 (83.3%)

n = Sample Size

Measures

The team of research assistants assigned to administer the battery of assessments consisted of 16 undergraduate and 2 graduate research assistants in the communication disorders program at Brigham Young University. Each blindly administered the complete assessment battery to all participants, unaware of individual language ability, and audio recording each testing session. The battery of assessments included in this study are the Narrative Language

Measure (NLM), a nonword repetition task (NWR), and a dynamic assessment of language (the DYMOND). The complete assessment battery was administered to students in their school in approximately 30-minute sessions in the Fall of 2018 and the Winter of 2019. Testing sessions took place over two days as needed to accommodate school schedules, potential fatigue, and individual participant circumstance.

CUBED: Narrative Language Measures (NLM)

The *NLM Listening* subtest of the *CUBED* (Petersen & Spencer, 2016) is a language assessment and progress monitoring tool and index measure that will be used to help establish language disorder and will also serve as a language sample for each participant. Each examiner read the assessment script and narrative, then prompted the student to retell the same narrative, providing on neutral encouragement (e.g., "just do your best"). All assessments were scored in real time by examiners and audio-recorded for further analysis. Results of the *NLM Listening* subtest, first-grade spring benchmark story 1 provide information for all students across all grades regarding their general oral language complexity including their ability to retell story grammar elements. Scores are reported out of a total of 34 points, are compared to national norms for grades K-3 and are referenced to sample-specific data for grades 4-6. A score at or below –1.5 standard deviations from the mean will serve to identify participants on this assessment.

Non-Word Repetition Task

Students completing the NRW task were asked to listen to an audio-recording and repeat 10, 2-6 syllable nonwords from the Children's Test of Nonword Repetition (CNRep; Gathercole et al., 1994) with two added, research-based nonwords. Additionally, this task will serve as an index measure to help establish language disorder in participants. Audio recorded responses were later scored according to the number of repeated syllables correct. Out of a possible 51 syllables, students scoring 36 (71%) or fewer syllables correct (Gathercole et al., 1994) will be identified as having a language disorder when they also meet the additional criteria listed above.

DYMOND Dynamic Assessment of Language

The DYMOND is a dynamic assessment of narrative language. This assessment includes 4 phases: a narrative retell pretest, a teaching phase, modifiability ratings completed by the examiner (see Figure 1), and a narrative retell posttest (Petersen et al., 2017). Each student was administered the English DYMOND in approximately 10 minutes, provided more or less time for individual responsiveness. Along with scoring the test in real time, examiners audio-recorded each session for additional analysis.

Dynamic Assessment Pretest. During the DYMOND pretest, the examiner reads a scripted narrative to the student and asks the student to retell the same narrative. The retell is scored in real time and is meant to measure a student's ability to use complex language (e.g., subordinating conjunctions such as when, because, and/or, after) and include story grammar elements in a dual-episode narrative (character, setting, problem, feeling, plan, attempt, consequence, feeling-2, plan-2, action-2, consequence-2, ending, and feeling). Students may score a maximum of 35 points on the pretest including their language complexity score (total of 9 points) and their story grammar subtotal (26 total points with 2 points per element).

Dynamic Assessment Teaching Phase. The purpose of the DYMOND teaching phase is to facilitate independent retelling of narratives and encourage student learning. During this phase, examiners model correct narrative retells including story grammar elements and complex language. Pictures and icons are used to visually support learning while the examiner retells the pretest narrative, offering direct instruction on each story grammar element and pointing to corresponding images for each (e.g., “Sam was disappointed- that is the feeling. How does Sam feel?”).

After direct instruction, the student is asked to retell the story using the pictures and icons, provided aid from the examiner when necessary to assist them in including all story grammar elements and appropriate language complexity. During direct instruction, examiners follow explicit teaching procedures outlined in the DYMOND including level 1 and level 2 prompt. A Level 1 prompt is used first when a student incorrectly retells or omits a story grammar element, and includes an open-ended question (e.g., “How did Sam feel?”). A Level 2 prompt is used when the student answers a level 1 prompt incorrectly and includes direct modeling of a correct response, asking the student to repeat it (e.g., “Sam felt disappointed. Now you say that.”). Examiners use an overcorrection procedure after either prompt, directing the student to back one-story grammar element and continue retelling the story, including the correct story element (e.g., “Awesome, you told me the feeling! Now start telling me the story again, this time starting at the problem. Remember to tell me how Sam was feeling.”) If the student is able to retell the narrative well without prompting, including most or all story grammar elements, the examiner may choose encourage use of complex language (e.g., targeting use of “because” or “after”). Students are then asked to retell the same narrative with only the story icons, with story images removed. Examiners may provide level 1 and level 2 prompting and the overcorrection procedure as needed.

Dynamic Assessment Modifiability. A modifiability score (24 possible points) and a final examiner judgment score (4 possible points) make up the DYMOND modifiability ratings (see Figure 1). Ratings are based on a 5-point Likert scale to rate the student across 6 areas of modifiability according to (1) their response to prompts, (2) the degree to which they transfer new knowledge, (3) their attention during the teaching phase, (4) how easy they are to teach, (5) their level of frustration, and (6) the degree to which their behavior causes disruption during the assessment (Peña, Gillam et al., 2006). Students are rated in each category (maximum of 4 points per category, with 4 points indicating considerable modifiability) with the sum of scores forming the modifiability score (total of 24 points). The examiner then completes a final examiner judgment score indicating the student's overall ease of learning observed to that point of the assessment (maximum of 4 points with 4 representing notable ease of learning and 0 representing considerable difficulty).

Dynamic Assessment Posttest. The DYMOND posttest requires the student to retell a different narrative than the story included in previous phases. The posttest story matches the pretest story in structure and difficulty such as in length, dual-episode structure, and use of complex vocabulary (e.g., tier-two words and subordinate clauses). Administration and scoring protocol for the posttest is equal to that of the pretest.

Test Administration: Fidelity and Inter-Rater Reliability

Fidelity

After receiving extensive training over several hours, four student team leaders were responsible for training each examiner in test administration and protocols. Research assistants were all communication disorders students and participated in subsequent a training session on how to administer each assessment included in the test battery. This hour-long training session required research assistants to practice administering the entire assessments battery to a team leader prior to administering tests to research participants. Once competency was established and research assistants demonstrated the ability to administer assessments in accordance with outlined test procedures with 100% accuracy, team members achieved approval from team leaders to administer assessments independently to participants. Team leaders remained onsite to provided general supervision and direction during testing sessions for new team members, to monitor fidelity, and ensure accuracy of assessment battery administration, including administration of the DYMOND.

Interrater Reliability

Interrater reliability was calculated based on rescored tests from randomly selected participants using a random number generator. Approximately 7% of students who were identified as typically developing and 12% of students identified as having a language disorder were included in this sample, examined by trained members of the research team who were blind to participants' language abilities. To analyze interrater reliability, independent examiners rescored pretest, modifiability ratings including final judgement and modifiability total scores, and posttest scores in real time from audio-recorded sessions corresponding to randomly selected participants. Rescored assessment results were compared with real-time results as scored by the initial test administrator to determine the percent and range of agreement.

Original interrater reliability of the pretest total scores was determined and then calculated based on a range of scores (between +/- 1-3 points respectively). Calculations are included in Table 3.

Table 3

Interrater Reliability Results

Score	IR1	Mod IR (range of points)
Pretest Total Score (35 total points)	31%	78-87.5% (+/- 2-3 points)
Modifiability Total Score (24 total points)	28%	75-88.0% (+/- 2-3 points)
Modifiability Judgment Score (4 total points)	60%	86-100.0% (+/- 1-2 points)
Posttest Total Score (35 total points)	25%	69-91.0% (+/- 2-3 points)

IR1 = Original Interrater Reliability Calculation; Mod IR = Modified Interrater Reliability Calculation based on the range of points identified

Statistical vs. Clinical Significance

Data obtained in this study were analyzed using one way between-subjects ANOVA to compare the effects of various demographic variables on total and final modifiability and posttest scores within groups of typically developing students and students identified as having a language disorder. Data were not compared across groups of students with and without language disorder. Given that data were not compared across groups, statistical differences between average test scores may be noted in the results of this study that may not have clinically significant implications.

Results

Typically Developing Students and Total Modifiability and Final Modifiability

Gender

A one way between-subjects ANOVA was conducted to compare the effects of gender of typically developing students on total modifiability. There was not a significant effect on total modifiability by gender, [F(1, 306) = 2.02, p = .16], $\eta_p^2 = .001$. A one way between-subjects ANOVA was conducted to compare the effects of gender on final modifiability. There was not a significant effect on final modifiability by gender, [F(1, 306) = 0.06, p = .81], $\eta_p^2 = .000$. Means and standard deviations for all outcomes are included in Table 4.

Grade

A one way between-subjects ANOVA was conducted to compare the effects of grade of typically developing students on total modifiability. There was a main effect on total modifiability by grade, [F(6,301) = 5.93, p = .000], $\eta_p^2 = .106$. Post hoc comparisons using the Tukey HSD test indicated that the mean total modifiability score for Kindergarten students was significantly different from the mean score of students in first-sixth grades, the mean score for first grade was significantly different from Kindergarten, fourth, fifth, and sixth grades, and the mean score for second grade was significantly different from Kindergarten, fourth, fifth, and sixth grades. The mean total modifiability score for second grade students was also significantly different from Kindergarten, fourth, fifth, and sixth grades.

A one way between-subjects ANOVA was conducted to compare the effects of grade of typically developing students on final modifiability. There was a main effect on final modifiability by grade, [F(6,301) = 4.59, p = .000], η_p^2 = .084. Post hoc comparisons using the Tukey HSD test indicated that the mean final modifiability score for Kindergarten students was significantly different from the mean score for second-sixth grade students, and the mean score for first graders was significantly different from students in second-sixth grades. The mean final modifiability score for second graders was also found to be significantly different from the mean score of fourth and fifth graders.

School Location

A one way between-subjects ANOVA was conducted to compare the effects of school location (rural or urban) of typically developing students on total modifiability. There was not a main effect on total modifiability by school location, [F(1, 306) = 0.59, p = .44], η_p^2 = .002. A one way between-subjects ANOVA was conducted to compare the effects of school location on final modifiability. There was not significant effect on modifiability by school location, [F(1, 306) = .09, p = .77], η_p^2 = .000.

English Language Proficiency and Dual Language Learner Status

A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency and DLL status of typically developing students on total modifiability. There was a significant effect on total modifiability by English language proficiency, [F(1, 306)

= 9.66, p = .002], η_p^2 = .031. A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency and DLL status of typically developing students on final modifiability. There was a significant effect on final modifiability by English language proficiency, [F(1, 306) = 11.78, p = .001], η_p^2 = .037.

Students With a Language Disorder and Total Modifiability and Final Modifiability

Gender

A one way between-subjects ANOVA was conducted to compare the effects of gender of students with a language disorder on total modifiability. There was not a significant effect on total modifiability by gender, [F(1, 52) = 1.11, p = .30], η_p^2 = .021. A one way between-subjects ANOVA was conducted to compare the effects of gender on final modifiability. There was not a significant effect on final modifiability by gender, [F(1, 52) = .01, p = .93], η_p^2 = .000.

Grade

A one way between-subjects ANOVA was conducted to compare the effects of grade of students with a language disorder on total and final modifiability. There was not a main effect on total modifiability by grade, [F(5, 48) = 1.55, p = .25], η_p^2 = .139. Similarly, there was not a main effect on final modifiability by grade, [F(5,48) = 1.38, p = .25], η_p^2 = .126.

School Location

A one way between-subjects ANOVA was conducted to compare the effects of school location (rural vs. urban) on total modifiability for students with a language disorder. There was not significant effect on total modifiability by school location, [F(1, 52) = 3.58, p = .06], ηp^2 = .064. A one way between-subjects ANOVA was conducted to compare the effects of school location (rural vs. urban) on final modifiability. There was not significant effect on modifiability by school location, [F(1, 52) = .654, p = .42], ηp^2 is .012.

English Language Proficiency and Dual Language Learner Status

A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency on total modifiability for students classified as having a language disorder. There was not a significant effect on modifiability by English language proficiency, [F(1, 52) = 2.22, p = .14], ηp^2 is .041. A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency on final modifiability with students who are classified as having a language disorder. There was not a significant effect on modifiability by English language proficiency, [F(1, 52) = .41, p = .52], η_p^2 = .008.

Typically Developing Students and Posttest Scores

Gender

A one way between-subjects ANOVA was conducted to compare the effects of posttest scores of typically developing students on total modifiability. There was not a significant effect on total modifiability by gender, [F(1, 306) = .01, p = .91], η_p^2 =.000.

Grade

A one way between-subjects ANOVA was conducted to compare the effects of grade of typically developing students on posttest scores. There was a main effect on posttest scores by grade, [F(6,301) = 10.18, p < .001], η_p^2 = .169. Post hoc comparisons using the Tukey HSD test indicated that the mean posttest score for first grade students was significantly different from the mean score of students in Kindergarten and second-sixth grades. The mean second grade score was also significantly different from mean scores of fifth and sixth graders.

School Location

A one way between-subjects ANOVA was conducted to compare the effects of school location (rural or urban) of typically developing students on posttest scores. There was not a main effect on posttest scores by school location, [F(1, 306) = 0.17, p = .68], η_p^2 = .001.

English Language Proficiency and Dual Language Learner Status

A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency (proficient or DLL) of typically developing students on posttest scores. There was a significant effect on posttest scores by English language proficiency, [F(1, 306) = 8.34, p = .004], ηp2 = .628.

Students With a Language Disorder and Posttest Scores

Gender

A one way between-subjects ANOVA was conducted to compare the effects of gender of students with a language disorder on posttest scores. There was not a significant effect on posttest scores by gender, [F(1, 52) = .511, p = .478], η_p^2 = .010.

Grade

A one way between-subjects ANOVA was conducted to compare the effects of grade of students with a language disorder on posttest scores. There was not a main effect on posttest scores by grade, [F(5,48) = 1.44, p = .227], η_p^2 = .131.

School Location

A one way between-subjects ANOVA was conducted to compare the effects of school location (rural vs. urban) on posttest scores for students with a language disorder. There was a significant effect on posttest scores by school location, [F(1, 52) = 9.91, p = .003], ηp^2 = .16.

English Language Proficiency and Dual Language Learner Status

A one way between-subjects ANOVA was conducted to compare the effects of English language proficiency on posttest scores for students classified as having a language disorder. There was a significant effect on posttest scores by English language proficiency, F(1, 52) = 7.07, p = .010], ηp^2 was .120.

Table 4

Modifiability and Posttest Scores of Typically Developing Students and Students With Language Disorder by Gender, Grade, School Location, and English Language Proficiency/DLL Status

		Typically Developing			Language Disorder		
Demographic and Modifiability Type	Variable	*n*	*M*	*SD*	*n*	*M*	*SD*
Gender							
Total Modifiability	Females	162	22.70	2.06	20	15.55	4.81
	Males	146	22.36	2.11	34	13.94	5.73
Final Modifiability	Females	162	3.78	0.44	20	2.10	0.91
	Males	146	3.77	0.42	34	2.07	1.05
Posttest	Females	162	16.13	4.43	20	9.65	4.39
	Males	146	16.08	3.96	34	8.65	5.29
Grade							
Total Modifiability	K	2	17.00	2.83	5	14.60	4.39
	1	25	21.48	2.24	16	12.25	5.90
	2	63	22.02	2.57	14	16.43	3.98
	3	47	22.45	1.86	5	18.40	3.44
	4	68	23.01	1.58	9	14.11	5.42
	5	62	22.89	1.78	5	13.40	7.92
	6	41	23.07	1.84	0	-	-
Final Modifiability	K	2	3.00	0.00	5	1.90	0.89
	1	25	3.48	0.51	16	1.69	0.95
	2	63	3.70	0.49	14	2.50	0.76
	3	47	3.79	0.39	5	2.60	1.14
	4	68	3.85	0.36	9	2.00	0.87
	5	62	3.85	0.36	5	2.00	1.58
	6	41	3.85	0.36	0	-	-
Posttest	K	2	13.00	8.49	5	8.60	3.91
	1	25	11.36	4.49	16	6.44	4.37
	2	63	14.83	3.80	14	10.29	4.34
	3	47	16.81	4.18	5	11.20	4.60
	4	68	16.71	3.30	9	10.22	6.12

	5	62	17.05	4.09	5	9.80	6.22
	6	41	17.88	3.60	0	-	-
School Location							
Total Modifiability	Rural	110	22.42	2.22	31	15.71	5.07
	Urban	198	22.61	2.01	23	12.96	5.57
Final Modifiability	Rural	110	3.76	0.45	31	2.18	0.97
	Urban	198	3.78	0.42	23	1.96	1.02
Posttest	Rural	110	15.97	3.99	31	10.71	4.28
	Urban	198	16.18	4.33	23	6.74	4.97
English Language Proficiency							
Total Modifiability	DLL	22	21.23	1.97	9	12.11	5.69
	Proficient	286	22.64	2.06	45	15.02	5.29
Final Modifiability	DLL	22	3.48	0.50	9	1.89	1.17
	Proficient	286	3.80	0.41	45	2.12	0.96
Posttest	DLL	22	13.64	4.18	9	5.22	4.68
	Proficient	286	16.29	4.16	45	9.78	4.69

n= Sample Size; M = Mean Score; SD = Standard Deviation; DLL = Dual Language Learner

Discussion

The purpose of this study was to examine the preliminary psychometric normative data for an English dynamic assessment of narrative language for a diverse group of elementary-age students with and without language disorder. Results obtained for this study partially confirm the hypotheses posed.

Typically Developing Students and Modifiability Scores

It was hypothesized that there would not be any significant differences in dynamic assessment modifiability between demographic subgroups (i.e., gender, English language proficiency or DLL status) within groups of children who were typically developing and children who had language disorder. The results of this study did not fully confirm our hypotheses. Although there were no significant differences in modifiability between gender (male/female) and school location (rural/urban) for all participants, the typically developing students had significantly different modifiability scores between younger students (kindergarten and first graders) and all other grades, and between students who were proficient in English and students who were identified as being DLLs.

Statistically vs. Clinically Significant Findings

Typically developing first grade students scored statistically significantly lower than other grades on final modifiability (*M* = 3.48). While mean scores for these students were found to be statistically different, in a clinical sense, the first-grade students still scored very close to the final modifiability ceiling score (4), and close to the mean modifiability scores for older grades (3.70-3.85). This idea begs the question whether the statistically significant difference is clinically significant, especially given that data obtained in this study were the results of within-group analysis. Scores of typically developing students were not compared to scores of students identified as having a language disorder, only to other typically developing students, students with preestablished adequate language abilities. In previous research using this same dynamic assessment, Petersen et al. (2017) found that students with language disorder consistently received scores of 0 and 1 on final modifiability and the majority of typically developing students received scores of 3 and 4. Thus, although the average score of the typically developing first grade students was statistically significantly different from their older peers, the first grade students represented in this group would still be considered typically developing given their final modifiability scores, and placed in the same category as their higher-grade peers. Therefore, when making classification judgments as to whether a child has a language disorder or not, the typically developing first grade students in this study would have been identified as typically developing given their modifiability scores on the DYMOND.

The significantly lower scores for first graders may be due to specific behaviors that are observed to complete the modifiability rating scale during the teaching phase of the DYMOND.

Examiners are directed to observe and rate 6 main areas of modifiability during the teaching phase including factors such as the child’s attention during the teaching phase, and the degree to which their behavior causes disruption during the assessment. These younger children may be exhibiting developmentally appropriate behavior for their age which may include decreased levels of attention and more distracting and disrupting movement or behavior during an assessment in comparison to their peers in older grades. With regard to clinical application, some behaviors may not be appropriate to observe in younger children because they are not necessarily indications of disorder. Another clinical implication to consider may include adjusting specific cut points for estimating sensitivity and specificity for younger students to allow for a wider variance in behavior with regard to the previously mentioned modifiability areas.

Considerations for Dual Language Learners

Students who were dual language learners (DLLs) were found to have statistically different mean modifiability scores in both total and final modifiability from students who were classified as being English proficient. Although their mean scores were statistically different, again, this may not be clinically significant enough to impact the interpretation of the results of this assessment to diagnose language disorder. DLL students who were typically developing scored close to the ceiling score on total and final modifiability, qualitatively similar to students who were classified as being English proficient (see table 2). The effect sizes (η_p^2 = .031 - .037) further confirm that although the findings are statistically significant, the meaningful application of these results is small. Although there may be some variance in modifiability ratings within groups of children who are typically developing, the research has shown that there is much greater variance between groups of children who are typically developing and those who have language disorder. In clinical use, dynamic assessment has shown to be an adequate measure of language ability within groups of culturally and linguistically diverse students, especially with regard to students who are DLLs. In their study on dynamic assessment of children who are English language learners, Kapantzoglou et al. (2012) found that a brief dynamic assessment of word learning was a "promising" diagnostic tool in differentiating English language learners who had a language disorder and those who were typically developing. Similarly, Ukrainetz et al. (2000) studied Arapahoe/Shoshone kindergarten students who were native English speakers but were identified as strong or weak language learners by teachers or researchers via classroom observation during instruction in Shoshone language as part of a cultural revival program at the school. They found that kindergarteners who were stronger language learners of a new language had higher modifiability scores on a dynamic assessment task teaching categorization than their age-matched peers who were classified as being weaker language learners.

Students With a Language Disorder and Modifiability Scores

It was hypothesized that there would not be any significant differences in dynamic assessment modifiability between demographic subgroups (i.e., gender, English language proficiency or DLL status) of children who had language disorder. With respect to this research question, the results of this study confirmed our hypothesis. When examining modifiability of students who were identified as having a language disorder, no significant differences were noted with regard to any demographic category. There was no main effect on total or final modifiability by gender, grade, school location, or English language proficiency. Peña, Gillam et al. (2006) found that modifiability scores were the strongest indicator of language impairment with regard to dynamic assessment scoring. Modifiability scores have been shown to have adequate sensitivity and specificity across grades (Petersen et al., 2020). Whether a student be in first or sixth grade, it is likely that their modifiability score will be low if they have a language disorder. Modifiability is not a measure of performance (e.g., number of correct story grammar elements), but a measure of how well a student can learn or adapt during the teaching phase of the assessment. This type of measurement is therefore unlikely to be impacted by grade or age, it is simply a measure of teachability regardless of demographic factors.

Typically Developing Students and Posttest Scores

It was hypothesized that there would not be any significant differences in posttest scores within the group of typically developing students across gender (male/female), and school location (rural/urban) demographics, but that there would be a significant difference between students of varying grades and those who were English proficient and students who were identified as being DLLs. Results from this study confirmed this hypothesis as well. When examining posttest scores of students who were typically developing, no significant differences were noted with regard to gender and school location. Significant differences were noted between grades and between groups of English proficient students and those who were DLLs. In their study on classification accuracy of narrative dynamic assessment, Peña, Gillam et al. (2006) found that among elementary aged students, there was no significant effect for gender on total story score which included similar measures to the posttest score represented in this study such as accurate retell of story grammar elements. Consistent with this finding, results of this study indicate that there was no significant effect on any score (i.e., total and final modifiability and posttest scores), by gender when examining results of students with language disorder and students who were typically developing.

As mentioned previously, posttest measures on a dynamic assessment are not the strongest indicator of disorder, especially with regard to culturally and linguistically diverse populations. Modifiability scores have been shown to be less biased than posttest scores on dynamic assessments of narrative language (Petersen et al., 2020). Students who are DLLs would therefore be expected to include fewer story grammar elements than their English proficient peers given their limited experience with English. However, posttest scores in conjunction with modifiability scores have been shown to increase the classification accuracy of a dynamic assessment (Peña, Gillam et al., 2006; Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020; Ukrainetz et al., 2000). Although there was a significant difference in posttest scores of students of various grades and varying exposure to English, when combined with a student's modifiability score, the posttest score can serve to identify disorder. Modifiability scores quantify a student's ability to learn language, and posttest scores demonstrate a student's ability to transfer information gained in the teaching phase such as improved language complexity or the inclusion of story grammar elements that were not represented in the student's performance on the pretest. Both scores taken into consideration offer a wider, clearer view of a student's performance and language ability. Even though DLL children across grades scored lower on the posttest than their older or more English-experienced peers, their relative performance compared to students with a language disorder is much higher. Children with a language disorder have much more difficulty transferring what they learn from the teaching phase to the posttest, including key story grammar elements, and using language complexity taught in the teaching phase (Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020). Thus, when using dynamic assessment to classifying disorder in clinical practice, it is best to consider both modifiability and posttest performance, especially with regard to students who are culturally and linguistically diverse.

Students With a Language Disorder and Posttest Scores

It was hypothesized that there would not be any significant differences in posttest scores within the group of students with language disorder across gender, and school location, but that there would be significant differences between students of varying grades, and between those who were English language proficient and DLL students. The results of this study did not fully confirm our hypothesis. When examining posttest scores of students with a language disorder, no significant differences were noted with regard to gender or grade. There were, however, significant differences in posttest scores of students with a language disorder from different school locations (i.e., rural vs. urban). Students with a language disorder from rural locations scored statistically significantly higher than their peers in urban locations. This could partially be due to other demographic features of these groups (e.g., the group of students from rural areas included a larger ratio of monolingual, English proficient students than the population of students from urban locations). We might expect the group of students from rural locations in this study to have a higher average score, even with regard to students with a language disorder, simply given that they were majority English proficient students.

All students who were identified as having a language disorder scored significantly lower than their typically developing peers on the posttest, however, as hypothesized, of those with a language disorder, students who were English language proficient had generally higher scores than those who came from linguistically diverse homes and were classified as being DLLs. As mentioned above, DLL students would be expected to include fewer story grammar elements and have a lower language complexity score than their English proficient peers on the posttest due to limited exposure to English. Petersen et al. (2020) found that within a group of culturally and linguistically diverse students, those with a disorder included fewer story grammar elements and lower language complexity scores on the posttest of the same dynamic assessment of narrative language used in the current study. Students who have a language disorder but are proficient in English would be expected to include more elements of a story with better language complexity than a child who has a disorder and limited exposure to linear narratives in the English language. Nevertheless, even children with a language disorder who are proficient in English consistently score considerably lower than their typically developing peers on posttest measures (see Table 4).

Limitations and Suggestions for Future Research

Within the group of typically developing students, kindergarteners scored statistically significantly lower on total modifiability (M = 17.00) than any other grade. It is possible to consider that kindergarteners had statistically significant, much lower scores than older grades due to greater variance in dynamic assessment performance. Examiners may have determined a wider range of modifiability scores for typically developing younger children (e.g., total modifiability SD for kindergarten was 2.83). It is important to note that only 2 kindergarten students without a language disorder were included in this study. This is an insufficient sample size to perform statistical inferencing, especially in comparison with older grades (1-6), each of which included a sample size of 25-68 students per grade. Further research should be conducted to include a larger sample size of younger students in order to adequately determine statistical and clinical significance of their performance on a narrative dynamic assessment of language. Likewise, this study included a very small sample size (*n* = 22) of students who were DLLs. In order to obtain an appropriate representation of typically developing, DLL students and their performance on a dynamic assessment of language, further research should be conducted to include a larger sample size of this demographic. Likewise, students who were identified as being DLLs in this study included majority native Spanish speakers. Research including a larger sample size of DLL students with a wider variety of native and home languages may be valuable in providing further evidence of the less-biased nature of the DYMOND with regard to classification of language disorder in culturally and linguistically diverse populations.

It should be noted that this study included relatively small sample sizes for every grade as far as number of students with a language disorder. Furthermore, no sixth-grade students with a language disorder were included in this study. Additional research should be conducted to include a larger sample of students with a language disorder to provide additional evidence regarding these students' performance on a dynamic assessment of language. Future research should consider further development of the preliminary psychometric, normative data obtained in this study.

Conclusion

The results of this study align with previous dynamic assessment research (Kapantzoglou et al., 2012; Peña, Gillam et al., 2006; Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020; Ukrainetz et al., 2000). Data indicate that although significant differences in modifiability and posttest performance were identified given various demographic factors, students with a language disorder consistently scored much lower on both modifiability measures and on the dynamic assessment posttest. This finding is consistent with previous research indicating that dynamic assessment modifiability scores have excellent sensitivity and specificity (Kapantzoglou et al., 2012; Kramer et al., 2009; Peña, 2000; Peña, Gillam et al., 2006; Peña et al., 2014; Petersen et al., 2017; Petersen et al., 2020; Roseberry & Connell, 1991; Ukrainetz et al., 2000). Thus, statistically significant data obtained in this study may not have clinically significant implications for identifying disorder, especially when modifiability and posttest scores are simultaneously considered in the diagnostic process. Furthermore, evidence provided in this study aligns with previous research indicating the less-biased nature of the DYMOND in classifying disorder within culturally and linguistically diverse populations (Petersen et al., 2017; Petersen et al., 2020). Upon the examination of various demographics (i.e., gender, grade, school location, and English proficiency/DLL status), the preliminary normative data identified in this study serve as a starting point for future research in the development of a norm-referenced dynamic assessment of narrative language.

References

American Psychological Association. (2020). *Norm-referenced test*. American Psychological Association Dictionary. https://dictionary.apa.org/norm-referenced-test

American Speech-Language-Hearing Association. (2019). *Demographic profile of ASHA members providing bilingual service, year-end 2019.* ASHA.org. https://www.asha.org/uploadedFiles/Demographic-Profile-Bilingual-Spanish-Service-Members.pdf

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.

Caesar, L. G., & Kohler, P. D. (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools, 38*(3), 190–200.

Exceptional Student Education Eligibility for Students with Language Impairment and Qualifications and Responsibilities for the Speech-Language Pathologists Providing Language Services, C.F.R. § 6A-6.030121 (2016). https://www.flrules.org/gateway/ruleNo.asp?id=6A-6.030121

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

Fulcher-Rood, K., Castilla-Earls, A. P., & Higginbotham, J. (2018). School-based speech-language pathologists' perspectives on diagnostic decision making. *American Journal of Speech-Language Pathology, 27*(2), 796-812.

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

Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. L*anguage, Speech, and Hearing Services in Schools, 27*(2), 109-121.

Individuals with Disabilities Education Act of 2004, 42 U.S.C. § 1247 *et seq*. (2004). https://www.govinfo.gov/content/pkg/USCODE-2011-title20/pdf/USCODE-2011-title20-chap33.pdf

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.

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.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*(4), 338-348.

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.

Owens, R., & Pavelko, S. (2017). Relationships among conversational language samples and norm-referenced test scores. *Clinical Archives of Communication Disorders, 2*(1), 34-50.

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

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. doi:10.1044/1092-4388(2006/074)

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech Language Pathology, 15*(3), 247-254.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220.

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. https://doi.org/10.1044/2016_JSLHR-L-15-0426

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

Petersen, D. B., Tonn, P., Spencer, T. D., & Foster, M. E. (2020). 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), 144-164.

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.

Selin, C. M., Rice, M. L., Girolamo, T., & Wang, C. J. (2019). Speech-language pathologists' clinical decision making for children with specific language impairment. *Language, speech, and hearing services in schools, 50*(2), 283-307.

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.

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*(6), 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*(2), 142-154.

United States Census Bureau (2018). *Languages spoken at home* (Table S1601) [Data set]. United States Department of Commerce. Retrieved from https://data.census.gov/cedsci/table?q=language&tid=ACSST1Y2018.S1601

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

APPENDIX A

Annotated Bibliography

American Psychological Association. (2020). *Norm-referenced test*. American Psychological Association Dictionary. https://dictionary.apa.org/norm-referenced-test

Objective: The purpose of this webpage was to provide the definition of a norm-referenced test.

Methods: The definition can be found alphabetically listed on the American Psychological Association's website.

Results: The definition provided indicates how norm-referenced tests are created and often utilized in clinical settings.

Relevance to Current Work: The definition of a norm-referenced test as provided by the APA serves to support rationale for the purpose of this study, specifically listing criteria that should be met in order to create a clinically useful NRT.

American Speech-Language-Hearing Association. (2019). *Demographic profile of ASHA members providing bilingual service, year-end 2019.* ASHA.org. https://www.asha.org/uploadedFiles/Demographic-Profile-Bilingual-Spanish-Service-Members.pdf

Objective: The purpose of this document was to provide public information regarding bilingual service delivery of SLPs certified through the American Speech-Language-Hearing Association (ASHA).

Methods: ASHA reported percentiles of certified members who meet criteria to provide bilingual service delivery per ASHA's bilingual clinical certification standards.

Results: ASHA reported that only 6.5% of SLPs with clinical certificate of competence met criteria to be classified as a bilingual service provider. It further reports that most of these SLPs report adequate skill in the provision of bilingual English-Spanish services.

Relevance to Current Work: These statistics provide evidence for the growing need for unbiased forms of assessment for CLD populations given that the vast majority of practicing SLPs do not indicate proficiency in the many diverse languages spoken by children in the United States.

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 examine various language disorder assessment tools and whether the psychometric properties of the assessments related to their frequency of clinical use.

Methods: Data for this were collected via a survey of 364 speech language pathologists regarding their clinical use of standardized assessment tools in diagnosing language impairment. Psychometric properties of 55 assessments were examined to determine the quality of each tool i.e., validity, reliability, etc.

Results: The results of this study indicate that practicing speech pathologists in elementary school settings preferred to use assessments that have a strong reputation in the field and are recommended by fellow professionals, regardless of empirical evidence of validity and reliability. The most commonly used assessment tools used in clinical diagnosis of language disorder included the Clinical Evaluation of Language

Fundamentals – Fourth Edition (CELF-4), the Preschool Language Scale, Fourth Edition (PLS-4), and the Peabody Picture Vocabulary Test (PPVT).

Relevance to Current Work: This study provides information regarding attitudes and preferences among speech pathologists in determining assessment tools for diagnosis of language disorder. It provides evidence to support standardization of assessment tools such as the DYMOND as standardized measures are perceived by many clinicians to be more valid forms of assessment.

Caesar, L. G., & Kohler, P. D. (2007). The state of school-based bilingual assessment: Actual practice versus recommended guidelines. *Language, Speech, and Hearing Services in Schools, 38*(3), 190–200.

Objective: The purpose of this study was to examine the frequency of use of formal, standardized assessment in comparison to the use of informal measure in the assessment of language ability in bilingual children. It also examined how often school-based speech pathologists adhere to recommended practices when assessing linguistically diverse populations.

Methods: Data were gathered via 130 surveys completed by school based SLPs in Michigan who were members of the Michigan Speech, Language, and Hearing Association, who indicated they were involved in the assessment of bilingual students. The survey included questions with regard to clinical background and practical experience, preparation and knowledge, characterization of diverse student populations and their ages/grades, listing of most commonly used assessment tools and procedures, and common assessment practices implemented in their evaluation of language skills.

Results: Researches found that school-based SLPs more commonly use standardized, norm-referenced assessment tools in the evaluation of language disorder of bilingual populations rather than informal measures I.e., DA.

Relevance to Current Work: This study provides evidence of current practices in the diagnosis of language disorder in culturally and linguistically diverse populations. It offers evidence to support the claim that SLPs often choose to utilize standardized, norm-referenced evaluation tools rather than informal measure to evaluate students. It provided evidence of the value of using DA in the diagnostic decision-making process to improve the ability to distinguish a language disorder from a language difference.

Exceptional Student Education Eligibility for Students with Language Impairment and Qualifications and Responsibilities for the Speech-Language Pathologists Providing Language Services, C.F.R. § 6A-6.030121 (2016). https://www.flrules.org/gateway/ruleNo.asp?id=6A-6.030121

Objective: The purpose of this regulation is to provide guidelines for the qualification of speech and language services for students in public schools within the state of Florida.

Methods: Various regulations are provided by the state of Florida in this document as a means of providing consistent standards for qualification of school-based SLP services across the state.

Results: Florida education regulations require at least one NRT score in order to qualify a student for school-based SLP services.

Relevance to Current Work: This document provides evidence that NRTs are widely considered to be necessary diagnostic tools by language and education professionals across the United States and are often required per state regulations in order to qualify students for services. Given the biased nature of many static NRTs in circulation, students may benefit from dynamic, less-biased forms of norm-referenced testing in order to accurately classify their language abilities and qualify them for school-based language services where appropriate.

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

Objective: The purpose of this resource was to outline theories of learning potential and adequate assessment of children with disabilities.

Methods: Feuerstein, Rand, and Hoffman assert that learning potential is a characteristic independent of demographic factors (i.e., home language, residential location, gender, ethnicity, etc.).

Results: Theories outlined in this resource indicate that measures such as modifiability, or a child's ability to learn, are not determined by cultural, social, or linguistic factors.

Relevance to Current Work: This resource provides support for the hypotheses stated in this study, specifically with respect to modifiability scores. Typically developing students would be expected to demonstrate similarly high levels of language learning ability as evidenced through their DYMOND modifiability scores (total and final) regardless of demographics such as gender, grade, school location, English proficiency or DLL status. Likewise, students with language disorder may be expected to demonstrate similarly low levels of language learning ability in comparison to their typically developing peers when modifiability is measured using a DA of language. Thus, the assertion provided in this resource (that the presence or absence of language disability should be the only factor to impact a student's true language learning potential) supports the use of modifiability measures in a DA to assess a student's language abilities.

Fulcher-Rood, K., Castilla-Earls, A. P., & Higginbotham, J. (2018). School-based speech-language pathologists' perspectives on diagnostic decision making. *American Journal of Speech-Language Pathology, 27*(2), 796-812.

Objective: The purpose of this study was to collect information from school based SLPs regarding common assessment practices for students with language disorder.

Methods: 39 SLPs from across the United States participated in structured phone interviews to investigate their assessment practices to identify language disorder in students. All participants were required to have a minimum of 5 years of experience working as a speech language pathologist and have assessment and treatment services for children with language disorder within their school. Interview questions included probes related to assessment tools and resources used, and the diagnostic decision-making process.

Results: Data collected for this study indicate that SLPs in schools use both formal and informal evaluation measures, however, place more weight in standardized testing performance and scores than on informal assessment results I.e., observation, teacher/parent reports, and language samples, when determining an appropriate diagnosis and classifying the severity of impairment of a student's language disorder.

Relevance to Current Work: This study provides evidence that SLPs place greater emphasis on standardized and norm-referenced testing results than on informal measures when classifying a child as having or not having a language disorder. Feedback from school based SLPs in this study indicate that clinicians would benefit from normative information reported for alternative forms of assessment such as a DA of narrative language. Reported norms for a DA may encourage SLPs to more frequently incorporate less-biased forms of testing in their assessment and diagnostic processes, specifically with regard to culturally and linguistically diverse student populations.

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

Objective: The purpose of this study was to present normative data with regard to the Children's Test of Nonword Repetition (CNRep).

Methods: This study included 600, 4-9-year-old children. Participants were administered the CNRep including the repetition of 40 nonwords.

Results: Results of this study were analyzed to create normative information for the CNRep. These data are to be used to assess children's language abilities and aid in the identification of language disorder.

Relevance to Current Work: The CNRep was utilized in the current study (referred to as the Nonword Repetition Task or the NWR) as a means of assessing language ability and classifying students as having or not having a language disorder. A score of 71% accuracy or lower on the NWR was used as one of three inclusionary criteria for participants to meet in order to be classified as having a disorder.

Hutchinson, T. A. (1996). What to look for in the technical manual: Twenty questions for users. L*anguage, Speech, and Hearing Services in Schools, 27*(2), 109-121.

Objective: The purpose of this study was to provide guidance for clinicians when interpreting results of standardized assessments. Hutchinson outlines psychometric principles of testing and encourages SLPs to investigate the adequacy of assessments before interpreting results or including them in the diagnostic decision-making process.

Methods: The researcher outlines several psychometric properties that should be considered before including any standardized assessment as a clinical diagnostic tool. He also provides questions to improve clinical and critical analysis of the appropriateness of each assessment with regard to the student or population being evaluated.

Results: The researcher answers a series of 20 questions that should be considered to adequately assess a test's psychometric integrity and diagnostic accuracy.

Relevance to Current Work: This study provides a rationale for the inclusion of assessment tools that have been shown to have adequate psychometric properties (i.e., sensitivity, specificity, reliability, and validity). It also provides rationale for the need to closely examine the demographic features of an assessment's normative sample before determining if the test will accurately assess and classify their language ability.

Individuals with Disabilities Education Act of 2004, 42 U.S.C. § 1247 *et seq*. (2004). https://www.govinfo.gov/content/pkg/USCODE-2011-title20/pdf/USCODE-2011-title20-chap33.pdf

Objective: The Individuals with Disabilities Education Act serves to provide free and appropriate public education, special education services, and resources to students with disabilities.

Methods: This law requires students to be assessed in a fair and appropriate manner and for relevant services to be provided for students with disabilities in order to help them achieve the highest level of independence and quality of life possible.

Results: IDEA requires SLPs to appropriately and adequately assess and provide intervention for students regardless of demographic factors (i.e., age, gender, religion, disability, school location, English proficiency/DLL status, ethnicity, etc.).

Relevance to Current Work: This law provides evidence of the importance of creating and using unbiased forms of assessment for CLD students in order to correctly identify language disorder in these populations. Dynamic assessments such as the DYMOND have been shown to have excellent classification accuracy of language disorder in populations such as these, where students may be DLLs or have insufficient exposure to cultural or linguistic aspects of the dominant culture of a student's residential area.

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.

Objective: The purpose of this study was to determine if a DA of word learning was an adequate measure to determine language disorder in bilingual children.

Methods: This study included 4- and 5-year-old participants who were native Spanish-speakers, 15 typically developing children, and 13 with language disorder. Each was administered a Spanish DA of word learning with a pretest-teach-posttest format.

Results: Data collected for this study indicate that students who were typically developing demonstrated better and quicker phonological-semantic association skills when introduced to new words than their peers with language disorder. This DA of word learning along with the Learning Strategies Checklist were shown to adequate measures for identifying language disorder in bilingual children.

Relevance to Current Work: This study provides evidence that DA is an adequate evaluation tool in the diagnosis of language disorder in linguistically diverse children, and that it can adequately identify children who are and are not disorder within a group of students who have similarly diverse language background and may be identified as DLLs.

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

Objective: The purpose of this study was to determine the diagnostic accuracy of a DA in differentiating language disorder from language difference in a culturally and linguistically diverse population. This study specifically investigated the test's diagnostic accuracy and cultural sensitivity based on the modifiability measure, and whether this measure was consistently adequate to identify disorder given various economic, cultural, and language proficiency and experience factors.

Methods: 17 third grade, First Nation students from Alberta, Canada participated in this study. Students were recruited from a school in the Samson Cree Nation Reserve. They were administered an English DA, with language disorder determined by a combination of teach, parent, and speech pathologist report and observation.

Results: Results of this study indicate that children with typical language ability demonstrated greater improvement and inclusion of assessment targets and improved transfer of learned skills across phases following direct instruction during the teaching phases than their peers identified as having a language disorder or delay. This DA was shown to be an adequate tool in diagnosing language disorder in a culturally diverse population.

Relevance to Current Work: Kramer et al., provide evidence that the modifiability measure in a DA is an adequate diagnostic measure for language disorder. This study further supports the clinical use of DA as a culturally sensitive, less-biased approach to the evaluation of language ability in diverse populations.

Laing, S. P., & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing services in schools, 34*(1), 44-55.

Objective: The purpose of this article was to discuss clinical issues related to the use of static NRTs and provide alternative, less-biased evaluation tools to use in the assessment language and literacy in CLD populations.

Methods: Laing and Kamhi present solutions to current clinical concerns with the use of static NRTs as a biased form of assessment with regard to CLD populations.

Results: Information presented in this article indicates that adequate assessment of CLD children should not be limited to static, standardized assessment, but rather include a variety of assessment procures. They present both DA and processing-dependent procedures are adequate alternative methods of assessment for these populations.

Relevance to Current Work: This article provides evidence that static NRTs are biased in many ways with regard to CLD populations. Laing and Kamhi discuss the biased nature of NTRs, specifically concerning testing content (i.e., stimuli, targets and methods), linguistic content (i.e., dialect, vocabulary used in the assessment), and inappropriate representation of CLD populations in normative samples. This article further supports the use of DA as an appropriate, less-biased form of assessment given that it seeks to identify disorder based on learning ability and not on static knowledge or a child's linguistic, social, or cultural experiences or practices.

McCauley, R. J., & Swisher, L. (1984). Use and misuse of norm-referenced tests in clinical assessment: A hypothetical case. *Journal of Speech and Hearing Disorders, 49*(4), 338-348.

Objective: The purpose of this study is to discuss common errors in the clinical use of NRTs and provide possible solutions and alternative means of assessment.

Methods: McCauley and Swisher illustrate examples of common misuse of NRTs in clinical practice. Hypothetical case studies are used in this study to illustrate common errors in the use of NRTs and why these errors may be causing the under- and over-identification of students with language disorder.

Results: Information provided in this article demonstrate several issues in the clinical use of NRTs, specifically the use of age-equivalent scores as a representation of a student's current performance, assuming a student's performance on certain test items indicates deficit within those areas of language, and the use of repeated assessments as a means of measuring progress.

Relevance to Current Work: This article discusses several reasons for the careful use of NRTs in clinical practice and supports the use of alternative forms of assessment to identify disorder and classify a student's current level of functioning with regard to language ability.

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.

Objective: The purpose of this study was to provide a systematic review of the diagnostic accuracy of a DA for language impairment in bilingual children and to examine their current clinical use.

Methods: Seven studies were reviewed using a meta-analysis procedure. 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 DA studies examined in this systematic review demonstrated higher scores on language assessments for typically developing (TD) participants than those with language impairment. Modifiability scores during the teaching phase consistently yielded higher scores for TD participants.

Relevance to Current Work: This study provides evidence that DA yield good classification accuracy (i.e., sensitivity and specificity) 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.

Owens, R., & Pavelko, S. (2017). Relationships among conversational language samples and norm-referenced test scores. *Clinical Archives of Communication Disorders, 2*(1), 34-50.

Objective: The purpose of this study was to compare the results of language sample analysis with standardized NRT results to evaluate the effectiveness of standardized testing in classifying language ability.

Methods: 16 children ages 3-7 years old participated in standardized testing and completed a 15-minute conversational language sample to collect data on mean length of utterance, total number of words, and words and clauses per sentence.

Results: Data collected for this study indicate that mean length of utterance, total number of words, and words per sentence values were shown to have a strong correlation with NRT results for individual students.

Relevance to Current Work: This article provides a discussion of common clinical practice with regard to static, standardized, NRTs. It also provides evidence to support the inclusion of alternate forms of assessment such as language sample analysis in the diagnostic decision-making process.

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

Objective: The purpose of this study was to investigate the components of modifiability and clinical implications related to modifiability measurements and their use in diagnosing language disorder.

Methods: This article reviews two studies, both of which employed a scaling system to quantify aspects of a student's modifiability.

Results: Results of this study indicate that modifiability measurements serve as an adequate classification tool in distinguishing students who are typically developing and those with low language abilities within CLD populations.

Relevance to Current Work: This study provides evidence that modifiability is an adequate, promising, less-biased measurement that should be included in the diagnostic process when classifying a student's language ability, particularly with regard to CLD groups.

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. doi:10.1044/1092-4388(2006/074)

Objective: The purpose of this study was to examine the reliability and classification accuracy of a narrative-based DA task.

Methods: This study included 2 experiments. The first experiment included 58 first and second grade children who were asked to tell 2 stories given wordless picture books. The second experiment included 71 first and second grade participants who were administered a DA (pretest, teaching phase, posttest). One group of typically developing (TD) and language impaired (LI) students received teaching, with a control group made up of TD children.

Results: Data collected for the first experiment in this study indicate that narrative measure had good internal consistency. Experiment two results indicate that participants who were TD and had received mediated learning had higher pretest to posttest scores than the control group and the group of students with LI. High sensitivity and specificity were found for this DA when modifiability scores were considered in conjunction with posttest scores.

Relevance to Current Work: This study provides evidence that narrative-based stimulus materials have good internal consistency and can be used for DA. It further supports the use of DA of narrative language to accurately identify children with LI.

Peña, E. D., Spaulding, T. J., & Plante, E. (2006). The composition of normative groups and diagnostic decision making: Shooting ourselves in the foot. *American Journal of Speech Language Pathology, 15*(3), 247-254.

Objective: The purpose of this study was to investigate and discuss what populations should be included in the makeup of a normative sample with regard to the objectives of individual normative assessments.

Methods: Researchers discussed psychometric and clinical implications of the normative sample including the impact on an assessment's classification accuracy. 32 assessments of child language were examined on the basis of their normative sample demographics and reported norms. A testing simulation was also conducted including one normative assessment which involved a normative sample of students with and without language disorder, and a second normative assessment which included a normative sample of only typically developing students.

Results: Peña et al., found that including a mixed group of students with and without language disorder may decrease an assessment's classification accuracy. Assessments that included mixed groups in their normative sample demonstrated lower group average scores and larger standard deviations, resulting in diminished diagnostic accuracy.

Relevance to Current Work: This study provides important discussion points with regard to normative assessment and the makeup of an adequate normative sample. Although researchers in this study found that assessments may have improved classification accuracy when only typically developing children are included in the normative sample, this study demonstrated the importance of considering specific demographic factors in the design of a normative sample to adequately reflect an assessment's objectives and outcomes and to serve as a basis for the appropriate interpretation of a student's performance on that assessment.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220.

Objective: The purpose of this study was to investigate the classification accuracy of an English DA of narrative language for children with limited English proficiency, classified as DLLs.

Methods: 54 Spanish-English bilingual participants were included in this study. One group consisted of 18 children with language disorder, with a control group including 18 typically developing, age- and gender- matched peers, and another group of 18 age-matched peers. Children participated in a variety of formal and informal assessment procedures, including a DA of narrative language.

Results: Data collected for this study indicate that a combination of modifiability ratings and posttest scores (i.e., posttest story grammar and language complexity/structure scores) yielded good sensitivity and specificity for the narrative language DA.

Relevance to Current Work: This study provides evidence that DA is a valuable tool in the classification of language ability of CLD and DLL children. Compared to their typically developing, age- and English proficiency- matched peers, children with a language disorder in this study consistently performed markedly lower on the DA with regard to modifiability and posttest measures. This study provides strong evidence that within a group of DLL students, a DA of narrative language demonstrates good sensitivity and specificity, and thus can be considered an appropriate and useful diagnostic tool in the classification of language ability of CLD and DLL populations.

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. https://doi.org/10.1044/2016_JSLHR-L-15-0426

Objective: The purpose of this study was to examine the diagnostic accuracy of an English DA of narrative language, specifically when used to identify language disorder in bilingual student populations.

Methods: This study included 42 kindergarten-third grade students who were classified as Spanish-English bilinguals. Each student participated in two sessions of DA testing consistent of a pretest-teach-posttest format.

Results: Data collected for this study indicate that a DA of narrative language yielded good sensitivity and specificity when modifiability scores, posttest measures, and duration of teaching phase were considered. Overall modifiability ratings were shown to have the best classification accuracy (i.e., 100% sensitivity, 88% specificity) of any one DA measurement.

Relevance to Current Work: This study utilized the same DA used in the current study and provides evidence that this the DYMOND is a valuable and appropriate tool to use in the diagnosis of language disorder in bilingual populations. It supports the use of modifiability and posttest scores to adequately classify a student's language ability and supports implementation of this assessment into regular clinical practice. It further provides cutoff scores for the DYMOND.

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

Objective: The purpose of this study was to investigate current research and methods used in narrative language interventions for preschool-aged students, specifically in connection with state academic and language standards.

Methods: This study provides a review of narrative-based language interventions for CLD preschool populations.

Results: Researchers found that narrative retell intervention in CLD preschool populations resulted in preschooler's ability to utilize story grammar elements and increased language complexity.

Relevance to Current Work: This research provides evidence to support the use of narrative-based language assessment and treatment in the management of language disorder in culturally and linguistically diverse populations.

Petersen, D. B., Tonn, P., Spencer, T. D., & Foster, M. E. (2020). 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), 144-164.

Objective: The purpose of this study was to investigate the diagnostic accuracy of a DA of inferential word learning in comparison with that of a static, standardized vocabulary assessment in the evaluation of Spanish-English bilingual students with a language disorder.

Methods: This study included 31 bilingual English-Spanish students ages 5-9 with typically developing language and language disorder. Each student was administered a DA of inferential word learning and a static vocabulary assessment.

Results: Data collected for this study indicate that the combination of modifiability and posttest scores of the DA of inferential word learning demonstrated better classification accuracy (i.e., sensitivity and specificity between 90-100%) than the static assessment of vocabulary.

Relevance to Current Work: This study provides evidence that DA may be a more appropriate and accurate form of evaluation than static, standardized assessments that are commonly used in clinical practice in the process of identifying language disorder in CLD students. It supports the excellent classification accuracy and psychometric integrity of DA identified in previous studies and promotes the clinical implementation of DA as a valid assessment tool for bilingual and CLD students.

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.

Objective: The purpose of this study was to determine whether children with language disorder would perform differently on language-learning tasks than their typically developing peers within a group of DLL students.

Methods: 26 Hispanic 4–6-year-old DLL children with emerging English proficiency were taught an invented morpheme in 2 separate groups; 1 group of students with language disorder and the other consisting of students who were identified as typically developing. Researchers were blind to group identity and language abilities prior to the experiment.

Results: Data collected for this study indicate that the group of students with typically developing language learned the invented morpheme more quickly and demonstrated statistically significantly higher scores on experimental assessment procedures than their language-disordered peer group.

Relevance to Current Work: This study provides evidence that a symptom of language disorder may include difficulty learning new aspects of language. This ideology supports the use of modifiability ratings following the teaching phase of a DA where the examiner rates the student based on ease of teaching, transfer of knowledge across steps, and other language-learning and behavioral measures. This study supports the assertion that students with a language disorder will demonstrated greater difficulty and require more time during the teaching phase of a DA to learn new aspects of language (i.e., new story grammar elements or vocabulary).

Selin, C. M., Rice, M. L., Girolamo, T., & Wang, C. J. (2019). Speech-language pathologists' clinical decision making for children with specific language impairment. *Language, speech, and hearing services in schools, 50*(2), 283-307.

Objective: The purpose of this study was to investigate how SLPs use research to inform clinical practice in everyday situations, specifically with regard to the assessment and intervention of students with language disorder.

Methods: 563 SLPs completed a web-based survey concerning clinical evaluation and treatment for language disorder. They responded to questions concerning various brief, diverse descriptions of invented sample students with language disorder.

Results: Data collected for this study indicate that a variety of methods are commonly used in the assessment and treatment of language disorder across clinical practice. Many SLPs indicated change in method with regard to specific characteristics of a given invented student. Most SLPs indicated standardized testing as an assessment recommendation in the identification of language disorder.

Relevance to Current Work: This study provides evidence that SLPs often rely on NRT scores to identify language disorder and qualify students for direct, skilled language services. It provides evidence of current state regulations promoting the use of NRTs to qualify students for special education services in schools (i.e., Missouri requires a student to score in the disordered range on a minimum of 2 NRTs to qualify for school based SLP services). Common practices in combination with current state requirements regarding the qualification for language services in public schools provide a rationale for the norming of a DA to improve clinical use of these less-biased forms of assessment and increase classification accuracy of language disorder in school settings.

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.

Objective: The purpose of this study was to investigate whether children with language disorder consistently score low and are adequately identified as having a disorder following administration of commonly used standardized language assessments.

Methods: Researchers examined test manuals of 43 commercially available static language assessments to investigate evidence of adequate classification accuracy of language disorder in pediatric populations.

Results: Data collected for this study indicate that children with language did not consistently score low on all static standardized language assessments. Furthermore, key psychometric information (i.e., sensitivity and specificity) were not reported in the majority of NRTs examined.

Relevance to Current Work: This study provides evidence that NRTs in general demonstrate inconsistent psychometric properties including adequate sensitivity and specificity and may not provide an accurate reflection of student's true language ability. Clinical implications of the results of this study support the use of alternate forms of assessment such as DA in conjunction with static NRT scores in the diagnostic process to improve classification accuracy of students. This study further supports the norming of assessments that have good sensitivity and specificity (i.e., the DYMOND) in order to improve classification accuracy in clinical settings when NRTs are required to identify disorder and qualify a child for services.

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*(6), 1245-1260.

Objective: The purpose of this study was to examine the prevalence of language disorder in young children.

Methods: Tomblin et al. investigated the prevalence of language disorder in a sample of 7,218 monolingual, English-speaking kindergarteners from rural, urban, and suburban areas in the upper Midwest region of the United States.

Results: Data collected for this study indicate that roughly 7.4% of children have language disorder and that it is a disorder that should be normally distributed within the general population. Participants were screened, following which those who failed the language screen were administered a diagnostic battery of assessments to identify students who had language disorder.

Relevance to Current Work: This study provides evidence that language disorder is a normally distributed disorder occurring in the general population, and that English proficiency (proficient/DLL) and school location (rural/urban) demographics are not statistically significantly correlated with higher incidence of language disorder.

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*(2), 142-154.

Objective: The purpose of this study was to examine the diagnostic accuracy of DA in Native American children and to determine whether it is a less-biased form of assessment within this CLD population.

Methods: 23 native English-speaking Arapahoe/Shoshone kindergarten students participated in a brief DA. Modifiability and posttest scores from the DA were compared with each student's performance on selected subtests of a standardized assessments.

Results: Researchers found that kindergarteners who were previously identified as weaker language learners of a new language had lower modifiability scores on the DA than their age-matched, stronger language learning peers. Modifiability scores were found to be an adequate differentiating measure in the classification of language learning ability (i.e., weak or strong language learning skills).

Relevance to Current Work: This study provides evidence that DA is an adequate tool in the characterization of language disorder in CLD student populations. Furthermore, it supports the use of modifiability and posttest measures as statistically adequate measures of disorder, as the DA in this study was shown to have good classification accuracy of language disorder within this population.

United States Census Bureau (2018). *Languages spoken at home* (Table S1601) [Data set]. United States Department of Commerce. Retrieved from https://data.census.gov/cedsci/table?q=language&tid=ACSST1Y2018.S1601

Objective: This data set lists the languages spoken in U.S. homes and the coordinating percentage of the population that speaks each language, based on reports collected in the 2010 census.

Methods: Projections were made for the year 2018 based on data collected via the 2010 census.

Results: Projections indicate that 21.9% of households speak a language other than English in the United States.

Relevance to Current Work: These data reflect the growing cultural and linguistic diversity found in the United States. It provides strong evidence to support the need for unbiased, culturally and linguistically sensitive forms of assessment that may be used clinically to improve classification accuracy of language disorder in increasingly diverse student populations in public schools.

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

Objective: This book contains a collection of Vygotsky's essays outlining psychological theories of cognitive development.

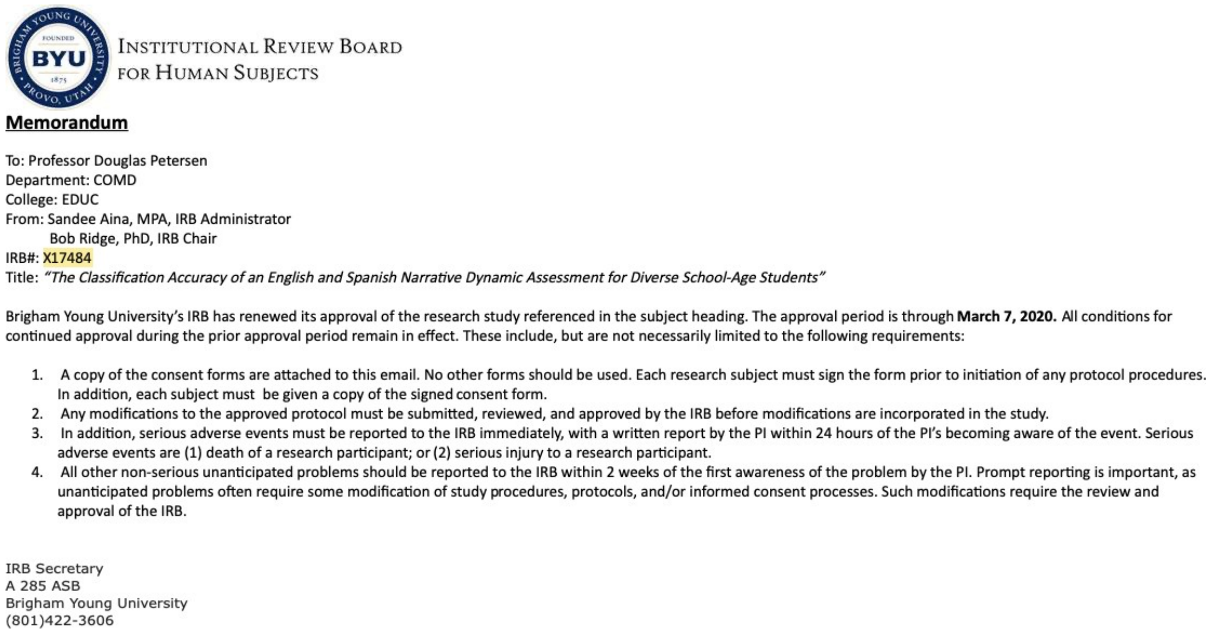
Methods: The purpose of these essays was to share early theories related to psychological and cognitive development in children.

Results: Vygotsky outlines several theories related to the early development of cognition and language including the principle of the zone of proximal development.

Relevance to Current Work: Vygotsky's work, specifically with reference to his theory of the zone of proximal development forms the basis of much of what we understand with regard to early language and psychological development. Furthermore, DA was created with this principle in mind, that children demonstrate learning capacity when given the opportunity to complete tasks just beyond their scope of abilities, provided an appropriate level of scaffolding. This is what the teaching phase of a DA aims to accomplish in order to correctly identify a student's current language abilities and their capacity to learn new language skills. The accurate measurement of a student's language-learning capability leads to good classification accuracy with regard to a diagnosis of typically developing or disordered language.

APPENDIX B

IRB Approval



INSTITUTIONAL REVIEW BOARD
FOR HUMAN SUBJECTS

Memorandum

To: Professor Douglas Petersen
Department: COMD
College: EDUC
From: Sandee Aina, MPA, IRB Administrator
Bob Ridge, PhD, IRB Chair
IRB#: X17484
Title: *"The Classification Accuracy of an English and Spanish Narrative Dynamic Assessment for Diverse School-Age Students"*

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

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

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