Diagnostic Accuracy of Nonword Repetition Tasks for the Clinical Assessment of Spanish-English Dual Language Learners: A Preliminary Investigation

Audrey Czirr
Brigham Young University
Diagnostic Accuracy of Nonword Repetition Tasks for the Clinical Assessment of Spanish-English Dual Language Learners:
A Preliminary Investigation

Audrey Czirr

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

Connie Summers, Chair
Kathryn Cabbage
Erika Feinauer

Department of Communication Disorders
Brigham Young University

Copyright © 2022 Audrey Czirr
All Rights Reserved
ABSTRACT

Diagnostic Accuracy of Nonword Repetition Tasks for the Clinical Assessment of Spanish-English Dual Language Learners: A Preliminary Investigation

Audrey Czirr
Department of Communication Disorders, BYU
Master of Science

Nonword repetition (NWR) has demonstrated significant potential as a less-biased language assessment measure for dual language learners (DLLs). However, there are currently no available guidelines for the use of NWR in a clinical setting. The purpose of this preliminary study is to develop initial recommendations for the clinical use of NWR tasks by determining the diagnostic accuracy and optimal cut-off scores for two NWR tasks and scoring methods, and to evaluate the clinical feasibility of NWR as an assessment measure. Participants included 23 DLL students with and without language disorder between the ages of 6 and 8. Spanish and English NWR tasks were administered in school classrooms and scored by percent phonemes correct (PPC) and number of whole words correct. Optimal cut-off scores resulting in the best sensitivity and specificity were calculated for each task and scoring method. Diagnostic accuracy was then compared for each task, combination of tasks, and scoring method. English PPC, Spanish PPC, and combined whole word scores yielded acceptable levels of sensitivity and specificity. Combined PPC scores resulted in excellent specificity, but inadequate sensitivity. Whole word scores for the tasks individually did not approach acceptable diagnostic accuracy. The current findings suggest that NWR can be feasibly implemented in the clinical setting and yield accurate results. English-Spanish whole word scores show potential as an accurate assessment measure for DLL children but should be investigated further. English-Spanish PPC scores appear to be appropriate for ruling out a language disorder, but are insufficient for ruling one in. These results provide preliminary support for the use of NWR tasks in the clinical assessment of DLLs as well as initial recommendations for their administration and interpretation.

Keywords: nonword repetition, dual language learners, language disorder, evaluation, percent phonemes correct, whole word
TABLE OF CONTENTS

TITLE PAGE ................................................................................................................................... i
ABSTRACT .................................................................................................................................... ii
TABLE OF CONTENTS ............................................................................................................... iii
LIST OF TABLES .......................................................................................................................... v
LIST OF FIGURES ....................................................................................................................... vi
DESCRIPTION OF THESIS STRUCTURE AND CONTENT .................................................. vii

Introduction ..................................................................................................................................... 1

   Current Practices and Challenges of Bilingual Language Assessment ......................... 2
   Means of Reducing Bias in Bilingual Language Assessment ............................................ 4
   Nonword Repetition Tasks .................................................................................................. 7
      Nonword Repetition in Monolingual Populations ......................................................... 8
      Nonword Repetition in Bilingual Populations .......................................................... 9
   Diagnostic Accuracy of Nonword Repetition .............................................................. 10
      Task Language .......................................................................................................... 11
      Cut-Off Scores ........................................................................................................... 11
      Scoring Methods ........................................................................................................ 12
   Purpose of the Study ....................................................................................................... 13

Method .......................................................................................................................................... 14

   Participants ...................................................................................................................... 14
   Measures ......................................................................................................................... 16
      Language Ability ....................................................................................................... 16
      Nonword Repetition ................................................................................................. 19
Procedures ............................................................................................................................... 20
Analysis ........................................................................................................................................ 22
Results ............................................................................................................................................. 23
Nonword Repetition Performance ............................................................................................ 23
Classification Accuracy of Individual Nonword Repetition Measures ...................................... 23
Cut-Off Scores ............................................................................................................................. 28
Classification Accuracy of Nonword Repetition Task Combinations ........................................ 29
Discussion ....................................................................................................................................... 32
Diagnostic Accuracy ...................................................................................................................... 32
Clinical Feasibility .......................................................................................................................... 35
Recommendations for Clinical Use ............................................................................................... 35
Limitations and Implications for Future Research ....................................................................... 37
Conclusion ...................................................................................................................................... 38
References ...................................................................................................................................... 40
APPENDIX A: Annotated Bibliography ......................................................................................... 48
APPENDIX B: Institutional Review Board Approval Letter ............................................................ 93
APPENDIX C: English Nonword Repetition Protocol ................................................................. 94
APPENDIX D: Spanish Nonword Repetition Protocol ................................................................ 95
APPENDIX E: Scoring Guidelines ................................................................................................. 97
LIST OF TABLES

Table 1  Participant Information ........................................................................................................16
Table 2  DYMOND Scores ......................................................................................................................19
Table 3  Mean Group NWR Scores ........................................................................................................23
Table 4  Diagnostic Accuracy and Optimal Cut-Off Scores For Individual NWR Measures ...24
Table 5  Diagnostic Accuracy of NWR Combinations ........................................................................31
### LIST OF FIGURES

| Figure 1 | DYMOND Clinical Decision Flowchart | 18 |
| Figure 2 | English Percent Phonemes Correct | 25 |
| Figure 3 | Spanish Percent Phonemes Correct | 26 |
| Figure 4 | English Whole Word | 27 |
| Figure 5 | Spanish Whole Word | 28 |
| Figure 6 | Combined Percent Phonemes Correct | 30 |
| Figure 7 | Combined Whole Word | 31 |
DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *Diagnostic Accuracy of Nonword Repetition Tasks for the Clinical Assessment of Spanish-English Dual Language Learners: A Preliminary Investigation*, is written in a hybrid format. The initial pages of this thesis are formatted to meet the traditional requirements for submission to the university, while the thesis report that follows is formatted according to journal publication conventions.

An annotated bibliography is included in Appendix A. Appendix B contains the Letter of Approval to Conduct Research from the Institutional Review Board. Protocols for the nonword repetition tasks examined in this study are included in Appendices C and D, followed by scoring guidelines in Appendix E.
Introduction

Children learning English as a second language represent approximately 10% of students in public schools in the United States, with Spanish reported as the home language for nearly 75% of these students (Hussar et al., 2020; Office of English Language Acquisition [OELA], 2020). The number of students learning English in the U.S. increased by more than 28% between 2000 and 2017, and this population is expected to continue to grow rapidly in coming years (OELA, 2020). Various terms have been used to describe these students including bilingual, limited English proficient (LEP), culturally and linguistically diverse (CLD), English Language Learner (ELL), and dual language learner (DLL). In the current study, students who are learning more than one language will be referred to as dual language learners (DLLs).

As the number of DLL students in the U.S. grows, school-based speech-language pathologists (SLPs) are tasked with the immense challenge of determining which students present with true language disorders, and which students demonstrate patterns typical of DLLs that change as their English proficiency increases. Distinguishing a language difference from a developmental language disorder (DLD) proves a difficult task, especially given that the linguistic patterns produced by DLLs often resemble those of monolingual children with language disorders (Barragan et al., 2018; Bedore & Peña, 2008; Gillam et al., 2013; Lazewnik et al., 2019; Levey et al., 2020). As a result, DLLs are more likely to be incorrectly identified as having a language disorder than monolingual children (Barragan et al., 2018; Gillam et al., 2013; Grimm & Schulz, 2014; Kraemer & Fabiano-Smith, 2017; Levey et al., 2020). For instance, a study conducted in Germany of kindergarteners acquiring German as a second language found the rate of misdiagnosis for the DLL children to be 27.3%, compared to only 14.5% for monolingual children (Grimm & Schulz, 2014). Similar patterns of overidentification for
Spanish-speaking students in the United States have also been found (Kraemer & Fabiano-Smith, 2017). Conversely, some SLPs exercise extreme caution in their clinical assessment of DLLs and inadvertently under-identify DLLs who present with a language disorder in their effort to mitigate the risk of overdiagnosis (Bedore & Peña, 2008; Kritikos, 2003; Levey et al., 2020).

Students who are misdiagnosed with a language disorder go on to receive speech-language services they do not need, while DLLs with language disorders who are not identified miss out on necessary services that could prevent significant academic difficulties later on (Kraemer & Fabiano-Smith, 2017; Levey et al., 2020). Thus, accurate assessment measures are essential to identify DLLs with language disorder and provide appropriate recommendations for intervention. A discussion of various challenges of bilingual language assessment is presented below, followed by a review of alternative testing methods and strategies that have been proposed to address these challenges, including nonword repetition.

**Current Practices and Challenges of Bilingual Language Assessment**

Past surveys of school-based speech-language pathologists in the U.S. have indicated that a majority of SLPs do not feel confident in their clinical assessment of DLLs (Guiberson & Atkins, 2012; Kimble, 2013). While SLPs report using different language measures in their assessment of DLLs, standardized language assessments in English remain the most commonly used assessment tool, with 85% of SLPs reporting using them often or always (Arias & Friberg, 2017; Dubasik & Valdivia, 2021; Gillam et al., 2013; Kraemer & Fabiano-Smith, 2017).

There are a number of limitations to standardized, norm-referenced language assessments for the purpose of identifying DLD. First, many assessments are not psychometrically sound and often fail to report on various methodological and/or psychometric properties, such as reliability, sensitivity, specificity, or likelihood ratios (Arias & Friberg, 2017; Denman et al., 2017;
McCauley & Swisher, 1984; Plante & Vance, 1994; Spaulding et al., 2006). There is also a question as to whether or not norm-referenced tests can accurately identify language disorder. In a review of 43 well-known language assessments, Spaulding et al. (2006) found that only five tests reached adequate levels of sensitivity and specificity to accurately discriminate children with language disorder from children with typical language. As a result of these limitations, SLPs are encouraged to exercise caution in their interpretation of standardized assessment results and to avoid relying solely on test scores to make diagnostic decisions.

In addition to the limitations just described, standardized language assessments are often biased against DLL children, making them inappropriate measures of language ability for this population (Bedore & Peña, 2008; Kraemer & Fabiano-Smith, 2017). The majority of standardized language assessments are knowledge-dependent, meaning that an individual’s performance is heavily influenced by their familiarity with the semantic content and grammatical structure of the test language (Campbell et al., 1997). Thus, children in the process of learning the test language are disadvantaged on such measures compared to those who have already acquired the language, regardless of their actual language learning ability (Bedore & Peña, 2008; Thordardottir & Brandeker, 2013).

Furthermore, the cut-off scores used to diagnose language disorder on many standardized assessments are derived from normative samples consisting primarily of monolingual speakers. DLLs differ from monolingual speakers in that their language knowledge is distributed across two languages; consequently, their performance can appear limited when only one language is considered (Peña et al., 2016). Given that many normative samples are not representative of DLL students, it is unlikely that the test norms will accurately reflect the performance of DLLs
(Bedore & Peña, 2008; Gillam et al., 2013). Such assessment measures risk overidentifying DLL students as a result.

To avoid the potential for misdiagnosis, the current standard for best practice in bilingual language assessment is to assess the DLL child’s language performance in each language (Levy et al., 2020; Peña et al., 2016). While this approach provides a more accurate representation of a child’s language abilities, it is often time-consuming and not always clinically practical, considering that the majority of SLPs in the U.S. are monolingual (American Speech-Language-Hearing Association [ASHA], 2021). Thus, efficient measures of language ability are needed for the assessment of bilingual populations.

**Means of Reducing Bias in Bilingual Language Assessment**

To address the concern of misidentification of DLLs, several alternative methods have been proposed as a means of reducing assessment bias for DLL children. Such methods include scoring modifications for monolingual language assessments, formal assessments normed on bilingual children, informal language measures, and processing-dependent measures. Clinicians are encouraged to incorporate several of these methods in their assessment of DLLs and to avoid making diagnostic decisions based on only one measure (Barragan et al., 2018; Bonifacci et al., 2020). In their meta-analysis of language assessment measures for Spanish-English bilingual children, Dollaghan and Horner (2011) did not find clear support for any single measure to discriminate language disorder from typical language. While many measures were found to be “suggestive” of DLD, they all required further evidence of language performance before a diagnosis could be confidently made (Dollaghan & Horner, 2011, p. 1086). Thus, best practice in bilingual language assessment involves synthesizing data from multiple assessment measures in
order to make culturally competent diagnostic decisions, described as a “converging evidence approach” (Barragan et al., 2018, p. 296; Castilla-Earls et al., 2020, p. 1125).

Some support has been found for the use of traditional monolingual language tests as an informative assessment measure for DLLs through the implementation of scoring modifications. For instance, bilingual or conceptual scoring methods, in which credit is awarded for a correct response in either of the child’s two languages, provide a more holistic view of the DLL child’s language ability (Levey et al., 2020; Peña et al., 2016). Additionally, alternative cut-off scores developed specifically for bilingual children have been suggested to increase the diagnostic accuracy of monolingual assessments. In their study of multiple English language assessments with Spanish-English bilingual children, Gillam et al. (2013) found poor specificity using the original monolingual cut-off scores. Alternative cut-off scores were found to improve diagnostic accuracy for bilingual children, though no single measure yielded both adequate sensitivity and specificity on its own.

Other informative approaches to bilingual language assessment include standardized assessments that have been normed on bilingual populations. For instance, some test developers have published Spanish versions of several mainstream English language assessments with bilingual norms, such as the Clinical Evaluation of Language Fundamentals – 4th Edition, Spanish (Wiig et al., 2006) and the Preschool Language Scales – 5th Edition, Spanish (Zimmerman et al., 2012). Recently, researchers have also developed assessments specifically for bilingual populations. While few are currently available, such assessments demonstrate good potential for accurately classifying the language ability of DLL children. In particular, the Bilingual English-Spanish Assessment (BESA) developed by Peña et al. (2014) has reported
Informal language assessment measures, such as language sampling and dynamic assessment, have similarly shown potential for reducing bias against DLL children. Language sampling and analysis can often provide a more accurate representation of a child’s language ability than standardized assessment tasks, as it allows clinicians to observe how a child uses language in natural contexts. Dynamic assessment describes how well a child is able to learn language, rather than how much they already know due to previous exposure. While both language sampling and dynamic assessment offer valuable information about a child’s language ability, recent surveys of practicing SLPs suggest they are implemented far less consistently than standardized language assessments in clinical practice. Arias and Friberg (2017), for instance, found that only 36% of SLPs collect language samples in the child’s native and second language often. Fewer than half of SLPs report including dynamic assessment in their evaluation of DLL students often or always (Arias & Friberg, 2017; Dubasik & Valdivia, 2021), with lack of time and training being commonly cited barriers (Arias & Friberg, 2017; Clark, 2019).

One recent effort to address these barriers is the development of the Dynamic Measure of Oral Narrative Language (DYMOND), which is a standardized, norm-referenced assessment of a child’s ability to produce academic language in the context of narrative retells (Petersen et al., 2017). The DYMOND can be administered in 15 minutes or less and involves real-time scoring to increase efficiency. It also establishes clear procedures for each testing and teaching phase, making it more accessible to clinicians. Perhaps more importantly, the DYMOND has reported good diagnostic accuracy with sensitivity and specificity at or above 80% for DLL students,
providing further evidence of its clinical utility (Clark, 2019; DeRobles, 2021; Frahm, 2021; Petersen et al., 2017).

Finally, processing-dependent measures have been reported to significantly minimize the effects of language experience by directly testing the processing mechanisms involved in language learning (Campbell et al., 1997). One example of a processing-dependent measure is nonword repetition. Thordardottir and Brandeker (2013) found that nonword repetition tasks were less affected by language exposure than traditional knowledge-dependent assessments. These measures may be an effective means of reducing bias in language assessment and therefore have the potential to increase the accuracy of diagnostic decisions for DLL students.

**Nonword Repetition Tasks**

Nonword repetition (NWR) is one processing-dependent measure of language ability that has been proposed as a method to reduce bias in bilingual language assessment. In an NWR task, nonsense words are presented to a child and the child is instructed to repeat each word. Generally, these nonwords are constructed to avoid word-likeness and predictability while still adhering to the phonotactic properties of the language, thereby minimizing the influence of language experience on task performance (Dollaghan & Campbell, 1998). NWR requires the listener to encode and temporarily store novel phonological representations, a process hypothesized to simulate word-learning (Gathercole, 2006; Summers et al., 2010). This process relies heavily on short-term phonological, or verbal, working memory, which is often cited as a hallmark deficit in children with DLD (Gathercole, 2006). Thus, poor performance on NWR may be indicative of such a deficit and imply the presence of a language disorder.
Nonword Repetition in Monolingual Populations

The current evidence supports the ability of NWR to accurately discriminate language disorder in monolingual and CLD populations (Dollaghan & Campbell, 1998; Kohnert et al., 2006). Dollaghan and Campbell (1998) were among the first to investigate the utility of NWR for diagnostic purposes. They developed a list of 16 English nonwords, designed to avoid lexical influence, ranging from one to four syllables. Significant differences in performance on the NWR task were observed between typically developing children and children with DLD. However, no differences in performance were found between White and African American participants, despite dialectal differences noted during conversation. Dollaghan and Campbell concluded that NWR was less biased against CLD populations and more accurately distinguished language ability than a traditional, norm-referenced language assessment. Similar findings were obtained by Kohnert et al. (2006) using the same nonwords, who observed significantly better performance by typically developing monolingual children than the language impaired group, particularly on nonwords with four syllables. They emphasized the need for NWR tasks to include nonwords with four or more syllables in order to be diagnostically useful.

Ebert et al. (2008) developed a list of Spanish nonword stimuli following the same development criteria used by Dollaghan and Campbell (1998). As expected, repetition accuracy of the nonwords was influenced by age and nonword length, with older children performing better than younger children, and performance generally declining for nonwords with three or more syllables. Importantly, no significant correlations were found between NWR accuracy and scores on the Preschool Language Scale-4 in either Spanish or English. This finding suggested that the NWR task was indeed a measure of language processing ability, rather than language proficiency. The same nonwords were later found to have adequate classification accuracy for

**Nonword Repetition in Bilingual Populations**

While studies investigating the use of NWR with bilingual children have suggested a high potential for diagnostic utility, they have also identified several factors that challenge the ability of NWR to identify language disorder in bilingual populations. First, typically developing DLLs and monolinguals perform differently on NWR tasks. Kohnert et al. (2006) observed that typically developing DLL performance fell somewhere between that of typically developing monolinguals and monolinguals with language disorder on the Dollaghan and Campbell (1998) NWR task. In addition to suggesting some level of bias inherent to the task, this pattern of performance makes it difficult to determine if a DLL child’s performance is typical or atypical. While Kohnert et al. (2006) determined that repetition accuracy was adequate to rule out the presence of a language disorder for DLL children, it was insufficient to rule it in. Therefore, the task was only somewhat informative.

Second, bilingual performance on NWR varies depending on the language of the task. When comparing the performance of Spanish-English DLLs on NWR tasks in both languages, Gutiérrez-Clellen and Simon-Cereijido (2010) observed inconsistent performance patterns among the children, with neither English nor Spanish appearing significantly more difficult than the other. While some children performed better on the task in their first or dominant language, others performed better in their second language. Furthermore, even some DLLs with typical language scored below the cut-off score used to identify DLD on the task in one language, but not the other.
Differences in language experience and exposure appear to contribute to this variability of performance, indicating that language knowledge may in fact influence performance on NWR to an extent (Windsor et al., 2010). Summers et al. (2010) found significant effects for length of English and Spanish exposure on the repetition accuracy of DLLs on NWR in each language. All children performed better on the Spanish task, but those with longer Spanish exposure repeated longer nonwords in each language more accurately than those with longer English exposure. A similar “practice effect” (Summers et al., 2010, p. 491) was later observed by Gibson et al. (2015) and hypothesized to result from the higher frequency of multisyllabic words in Spanish than English. Cumulative language experience was concluded to be a significant influencing factor on DLL performance on NWR (Gibson et al., 2015; Summers et al., 2010), leading Irizarry-Pérez et al. (2021) to suggest including longer nonwords in Spanish than in English for the tasks to be of comparable difficulty. Significant, if somewhat weaker, associations have also been reported between NWR performance and current levels of language use (i.e., input and output) for each language (Summers et al., 2010; Thordardottir & Brandeker, 2013). As Spanish-English DLLs as a group do not demonstrate consistent performance on NWR in either language, it is unlikely that administering NWR tasks in only one language will provide an accurate representation of their language ability. Thus, analysis of performance in both languages is necessary for NWR to be a clinically useful measure for DLL populations.

Diagnostic Accuracy of Nonword Repetition

Studies of NWR have reported various levels of diagnostic accuracy for the task. Diagnostic accuracy refers to the ability of an assessment measure to discriminate between typical and atypical performance and can be reported in multiple ways. Sensitivity represents the number of individuals who are accurately identified as having a disorder, while specificity
represents the number of individuals who are accurately identified as not having the disorder (Plante & Vance, 1994). A recent meta-analysis of 13 studies of NWR for the identification of language disorder in bilingual speakers of various languages found a significant range of sensitivity and specificity for different NWR tasks (Ortiz, 2021). There are several factors which may influence the ability of NWR to accurately discriminate DLD in bilingual children; these factors are discussed below.

**Task Language**

Multiple studies have found that the ability of NWR to accurately discriminate DLD from typical language in DLL children depends largely on the language of the task, though the findings of these studies have been mixed. Armon-Lotem and Meir (2016) and Kelly (2021) both identified NWR in the child’s first language (L1) as a more accurate measure to discriminate DLD in bilingual children than NWR in the second language (L2). In contrast, Windsor et al. (2010) reported sensitivity was much higher for NWR in English than in Spanish for bilingual children. Gutiérrez-Clellen and Simon-Cereijido (2010) found that neither task yielded adequate diagnostic accuracy when only one language was considered independent of the other. However, high levels of sensitivity and specificity were achieved when performance in both languages was considered together. Summers et al. (2010) similarly highlighted the importance of considering both languages in the assessment of DLL children and encouraged the use of NWR in both languages to account for differences in performance across languages.

**Cut-Off Scores**

The diagnostic accuracy of NWR tasks for DLL children is additionally influenced by the cut-off scores being used. Differences in performance between bilingual and monolingual speakers indicate that cut-off scores reflecting monolingual performance on NWR are an
inappropriate comparison standard for DLL children. In their investigation of Russian and Hebrew NWR tasks for Russian-Hebrew bilingual children, Armon-Lotem and Meir (2016) found the monolingual cut-off scores on both tasks had poor accuracy (57-78%) for identifying bilingual children with DLD. However, when bilingual cut-off scores were applied and performance in both languages was considered, overall accuracy increased to 89%. The greatest level of diagnostic accuracy was achieved when performance on NWR was considered together with other processing-dependent tasks using bilingual cut-off scores, underscoring the need for multiple measures of language performance in order to make accurate diagnoses.

Cut-off scores reflecting DLL performance are essential for NWR to be a valid measure of their language ability. To our knowledge, bilingual cut-off scores for Spanish-English speakers have not been specified for NWR tasks in either language. These must be established for NWR to be a clinically informative measure in the language assessment of Spanish-English DLLs.

**Scoring Methods**

Several methods of scoring NWR accuracy have been identified in the literature, with different levels of sensitivity and specificity reported for each method. In the percent phonemes correct (PPC) method, the child receives credit for every phoneme they repeat correctly for each nonword. Any substitutions or omissions of phonemes are scored as incorrect. Generally, phoneme additions or distortions are not counted as errors (Dollaghan & Campbell, 1998). PPC is calculated by dividing the number of phonemes the child produces correctly by the total number of phonemes included in the task. Whole word, or item-level scoring awards credit only for each nonword the child repeats accurately. Whole word scoring provides a more conservative assessment of the child’s performance as their total score reflects the number of nonwords they
produced correctly, leaving less room for error than PPC. This scoring method is also more straightforward and less time-consuming than PPC and may therefore lend itself more easily to clinical practice (Guiberson & Rodríguez, 2013; Roy & Chiat, 2004).

While both scoring methods (i.e., whole word and PPC) have been found to yield adequate levels of sensitivity and specificity, some studies have suggested that whole word scoring more accurately discriminates DLD than PPC. Guiberson and Rodriguez (2013) compared the classification accuracy of both scoring methods using a Spanish NWR task with Spanish-speaking preschool children with and without DLD. They found that PPC scoring resulted in adequate specificity, but poor sensitivity. In contrast, whole word scoring achieved adequate levels for both sensitivity and specificity, resulting in greater diagnostic accuracy overall.

The current study builds upon another work that similarly identified whole word scoring as having greater accuracy than PPC (Kelly, 2021). However, the findings differed from previous studies in that they suggested whole word scoring may not be as clinically useful as PPC scoring. While the difference in average scores between the DLD and typically developing (TD) groups was statistically significant, it was extremely small; the DLD group accurately repeated an average of 0.62 of 16 nonwords, while the TD group averaged only 2.08. The marginal difference in performance between the two groups obtained using whole word scoring would significantly limit the practicality of NWR as an assessment measure. Thus, further investigation of the diagnostic accuracy and clinical utility of both scoring methods is needed.

**Purpose of the Study**

Despite the promising ability of NWR to accurately identify language disorder in bilingual children, there currently exists little to no information to guide its use beyond research
purposes. In particular, the lack of clinical guidelines and specified cut-off scores to achieve the greatest diagnostic accuracy makes it difficult to extend the use of NWR to clinical practice. These must be established in order for NWR to be a clinically useful measure for accurate language assessment of DLL children. The current study constitutes a preliminary investigation as part of a larger work that seeks to determine such guidelines. Specifically, the following questions are examined:

1. Which NWR task or combination of tasks yields the greatest sensitivity and specificity for identifying language disorder in school-age Spanish-English DLLs?
2. Which scoring method (i.e., whole word or PPC) yields the greatest sensitivity and specificity?
3. Which cut-off scores yield the highest sensitivity and specificity?
4. Can NWR tasks be feasibly implemented as a language assessment measure in the clinical setting?

Based on findings from previous studies, we hypothesized that a combination of Spanish and English NWR tasks would result in the greatest diagnostic accuracy. Consistent with findings by Guiberson and Rodriguez (2013) and Kelly (2021), we also anticipated whole word scoring to discriminate language ability of DLL children more accurately than PPC. We expected NWR tasks to be clinically feasible given their short administration time and ease of scoring.

Method

Participants

This study consisted of a sample of 23 first and second grade, Spanish-English DLL students between the ages of 6 and 8. Data collection and analysis for this study were approved by the university Institutional Review Board. Participants were selected from a larger sample of
both monolingual and bilingual children from three elementary schools in the mountain west who were recruited as part of a broader, ongoing study of narrative language intervention (Petersen et al., 2022). Parents were informed of their child’s participation in this project as a district-wide initiative. Each student completed a child assent form indicating their willingness to participate.

Demographic information for each participant was collected from their individual schools and is represented in Table 1 below. All participants were classified as ELLs by their schools and were receiving associated services at the time of this study. Spanish was reported as the home language for all participants. Per school report, two participants had an Individualized Education Plan (IEP) for language impairment at the time of this study. A third participant was reported to have an IEP but was excluded from this study due to incomplete data regarding IEP classification. Eligibility for free school lunch was obtained as an indicator of socioeconomic status. Of the 23 participants, 18 (78%) qualified for free lunch, while five (22%) did not.

Participants were divided into two groups: those with DLD (n=11) and those with typically developing language (TD; n=12). A converging evidence approach, which considered three markers of language ability (i.e., eligibility for school speech-language services and performance on both a static and dynamic measure of language), was used to determine group placement (Castilla-Earls et al., 2020). For inclusion in the DLD group, participants were required to meet two of the following three criteria: (a) score below 1 SD below the mean on the Narrative Language Measures (NLM) Listening subtest of the CUBED assessment (Petersen & Spencer, 2016); (b) receive a classification of “language learning disorder” on the Dynamic Measure of Oral Narrative Discourse (DYMOND); (c) be receiving school-based speech-
language services for a language disorder at the time of the study. To be included in the TD group, participants may not have met two of the three criteria described above.

Table 1

*Participant Information*

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>Total participants (N=23)</th>
<th>DLD (n=11)</th>
<th>TD (n=12)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n</td>
<td>%</td>
<td>n</td>
</tr>
<tr>
<td>1st grade</td>
<td>18</td>
<td>78</td>
<td>9</td>
</tr>
<tr>
<td>2nd grade</td>
<td>5</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>Male</td>
<td>16</td>
<td>70</td>
<td>9</td>
</tr>
<tr>
<td>Female</td>
<td>7</td>
<td>30</td>
<td>2</td>
</tr>
<tr>
<td>IEP</td>
<td>2</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>Free lunch</td>
<td>18</td>
<td>78</td>
<td>10</td>
</tr>
</tbody>
</table>

*Note.* DLD = developmental language disorder. TD = typically developing. IEP = Individualized Education Plan.

**Measures**

**Language Ability**

Two language assessment measures, the NLM Listening subtest and the DYMOND, were used to inform group placement as either having DLD or typically developing language. Participants completed each assessment following the standard administration and scoring protocols described in the test manuals.

The NLM Listening is a brief, standardized measure that assesses complex, academic language skills through the production of narrative retells (Petersen & Spencer, 2012). It can be used as a stand-alone measure or in conjunction with additional subtests. For the purposes of this study, only the NLM Listening was administered. Participants were instructed to listen carefully as they heard a short narrative, after which they were asked to retell the same story. Scoring was completed in real time, with points awarded for story grammar and language complexity of the student’s retell. Raw scores for each retell were converted to standard scores and compared to the
bilingual Spanish-English norms provided in the test manual. Mean standard scores were unable to be calculated for each group, as 16 participants received a standard score of <64 (first grade) or <66 (second grade). However, all but one participant scored below 1 SD below the mean.

The DYMOND is a dynamic assessment of a child’s ability to learn language using a pretest-teach-posttest model (Petersen et al., 2017). After listening to a short narrative, participants were asked to tell the same story. They then participated in a brief teaching period regarding story grammar. Following the instruction, participants listened to a second narrative and retold the story to the examiner. Pre- and post-test retell scores were determined by the number of story grammar elements included and measures of language complexity. Total and final modifiability scores were obtained based on examiner ratings of the child’s responsiveness to teaching and transfer of targets (Petersen et al., 2017). Scores were compared to normative data to classify participants’ language ability as shown in Figure 1 below.
Figure 1

DYMOND Clinical Decision Flowchart

Group performance on the DYMOND is reflected in Table 2. The mean total posttest score for the TD group was 9.83. As expected, the DLD group scored significantly lower than the TD group, with a mean score of 4.18. The TD group additionally had a mean total modifiability score of 21.08 and a mean final modifiability of 3.75. The mean total modifiability and final modifiability scores for the DLD group were 14.05 and 2.10, respectively. Modiﬁability scores were unavailable for one participant in the DLD group.

**Table 2**

**DYMOND Scores**

<table>
<thead>
<tr>
<th>Measure</th>
<th>DLD</th>
<th>TD</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Total Pretest</td>
<td>3.82</td>
<td>2.560</td>
</tr>
<tr>
<td>Total Posttest</td>
<td>4.18*</td>
<td>3.920</td>
</tr>
<tr>
<td>Total Modiﬁability</td>
<td>14.05*</td>
<td>4.821</td>
</tr>
<tr>
<td>Final Modiﬁability</td>
<td>2.10*</td>
<td>1.075</td>
</tr>
</tbody>
</table>

*Note. DLD = developmental language disorder. TD = typically developing. *p < .05.*

**Nonword Repetition**

Participants completed two NWR tasks, one in English and one in Spanish. The English task consisted of a set of 16 nonwords adapted from Dollaghan and Campbell (1998) ranging from one to four syllables and constructed to avoid word-likeness. To reduce the potential for articulatory errors, they did not contain late-developing consonants or consonant clusters and only included tense vowels. The Spanish NWR task, adapted from Ebert et al. (2008), consisted of 20 nonwords between one and five syllables. These nonwords were constructed following the same standards described by Dollaghan and Campbell (1998). The current study followed the same guidelines for PPC scoring established in each of these studies.
Procedures

Participants completed all tasks in a 20-40 minute session at their elementary school. Sessions were conducted in a testing classroom with multiple examiners and participants in the room at a time. Tasks were administered by trained undergraduate or graduate research assistants. To ensure fidelity of task administration, research assistants completed task checklists throughout each session.

NWR tasks were introduced with the instructions, “We’re going to listen carefully to some made-up words. You’re going to listen to each word and then repeat it exactly like you hear it. Let’s practice.” Three practice nonwords were presented before the test items to confirm participant understanding of the task. If the child did not repeat a practice nonword, feedback was provided before the next item. If a participant responded with a similar-sounding real word, the research assistant reminded the child that the words were not real words. Following completion of the practice items, the participant heard the instructions, “Now we’re going to listen to more words. Some will be short and some will be long. Repeat the word exactly like you hear it. Listen carefully because each word will only be played once.”

To simulate a realistic clinical environment, nonwords were presented as an audio recording over a computer or phone speaker and were played once each. English nonwords were presented first, immediately followed by the Spanish nonwords. Participants responded after each nonword, with prompts to “repeat what you heard” as needed. Their responses were audio recorded and later transcribed for scoring by trained bilingual research assistants who were blinded to participants’ language ability.

Repetition accuracy was calculated using two different scoring methods. In the first scoring method, percent phonemes correct (PPC), participants were awarded credit for each
phoneme they accurately repeated per nonword. Following the scoring guidelines described by Dollaghan and Campbell (1998), substitutions and omissions were counted as incorrect, but distortions and additions were not. A scoring protocol adapted from Gray et al. (2019) and found in Appendix E provided more detailed guidelines for lining up phonemes in the child’s response against the target nonword to determine the accuracy of each phoneme. Vowels were matched with vowels and consonants with consonants, generally with the child’s response lined up from the beginning of the nonword. However, the rater was permitted to “slide” phonemes or syllables in the child’s response to line up with the most closely matching phonemes or syllables of the target nonword, provided that the overall syllable structure and order of phonemes were maintained. Added phonemes or syllables could also be “popped” out to maximize the child’s points. PPC was then determined by dividing the number of phonemes the child repeated correctly by the total number of phonemes for the task. Using the second method, whole word scoring, participants were awarded a score of 1 for each nonword they repeated accurately and a score of 0 for each nonword they repeated in error. Substitutions, omissions, and additions were counted as incorrect, but distortions were accepted as correct.

Thirty percent (7/23) of NWR tasks in each language were randomly selected for scoring by a second research assistant to determine interrater reliability for PPC scoring. Interrater reliability for the English task ranged from 82.3% to 95.8%, with an average of 89.12%. For the Spanish task, interrater reliability ranged from 92.5% to 99.2 %, with an average of 95.48%. Differences in scores between the two raters for these samples were reviewed and resolved by a third bilingual research assistant to obtain a final score.
Analysis

Differences in NWR performance between DLD and TD groups for each task were analyzed through independent sample t-tests, with group membership as the independent variable and repetition accuracy as the dependent variable. Receiver operator characteristic (ROC) analyses were conducted for each NWR measure individually, which calculated sensitivity and specificity values for all possible cutoff scores to identify the cutoff resulting in the most optimal sensitivity and specificity. ROC curves were generated by plotting sensitivity against the false positive rate (1-specificity) for each cutoff score, with area under the curve (AUC) computed as a measure of overall classification accuracy. Per recommendations for interpretation by Carter et al. (2016), an AUC at or above .90 was considered excellent, between .80 and .89 was considered good, and between .70 and .79 was considered fair. An AUC below .70 was considered non-useful.

To determine the classification accuracy of the Spanish and English tasks combined using the newly identified cutoff points for each task, two binary logistic regression analyses were performed for PPC and whole word scoring. Group classification served as the binary dependent variable, with English and Spanish NWR measures as the continuous predictor variables. The results of the binary logistic regressions were used to generate ROC curves for NWR task combinations to calculate the AUC. Sensitivity, specificity, and total classification accuracy were then calculated for each combination. Sensitivity and specificity levels at or above 90% were considered good, and levels between 80 and 89% were considered fair (Plante & Vance, 1994). Levels below 80% were considered unacceptable.
Results

Nonword Repetition Performance

Group performance for each NWR task and scoring method was analyzed to determine significant differences between groups. Mean scores for each group are reported in Table 3 below. The TD group performed significantly better than the DLD group on NWR in both English and Spanish using PPC scoring. The mean score on the English task using PPC scoring was 70.57 for the TD group and 50.16 for the DLD group. On the Spanish task, the mean PPC score was 80.90 for the TD group and 68.49 for the DLD group. The DLD group performed significantly lower than the TD group on the English task using whole word scoring, with a mean score of 2.00 compared to a mean of 4.75 words for the TD group. However, no statistically significant differences were found between groups on the Spanish task with whole word scoring.

Table 3

<table>
<thead>
<tr>
<th>Measure</th>
<th>DLD</th>
<th></th>
<th>TD</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>EPPC</td>
<td>50.16*</td>
<td>18.159</td>
<td>70.57*</td>
<td>10.862</td>
</tr>
<tr>
<td>SPPC</td>
<td>68.49*</td>
<td>12.285</td>
<td>80.90*</td>
<td>5.691</td>
</tr>
<tr>
<td>EWW</td>
<td>2.00*</td>
<td>1.342</td>
<td>4.75*</td>
<td>2.094</td>
</tr>
<tr>
<td>SWW</td>
<td>8.18</td>
<td>2.994</td>
<td>9.33</td>
<td>3.114</td>
</tr>
</tbody>
</table>

Note. TD = typically developing. EPPC = English percent phonemes correct. EWW = English whole word. SPPC = Spanish percent phonemes correct. SWW = Spanish whole word. *p < .05.

Classification Accuracy of Individual Nonword Repetition Measures

The first two research questions posed in this study concerned classification accuracy of different NWR tasks and scoring methods. ROC analyses were conducted to identify the sensitivity and specificity levels at the optimal cut point for PPC and whole word scoring in
Spanish and English, shown in Table 4. ROC curves for each task and scoring method are found in Figures 2-5, with area under the curve (AUC) representing overall accuracy.

Three of the four measures, English PPC, Spanish PPC, and English WW scores yielded good or excellent AUC by themselves. All four measures were shown to yield acceptable specificity; however, acceptable sensitivity was only achieved with PPC scoring. Spanish PPC yielded the best sensitivity and specificity overall, at 100% and 81.8%, respectively. English PPC was similarly found to have good, though somewhat lower, sensitivity at 91.7%. While Spanish whole word scores resulted in the best specificity (90.9%) of the four measures, both sensitivity and AUC were poor. Sensitivity increased to 75% for English whole word scores, though this still fell below the acceptable range.

Table 4

<table>
<thead>
<tr>
<th>Measure</th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Cut-off</th>
</tr>
</thead>
<tbody>
<tr>
<td>EPPC</td>
<td>0.905</td>
<td>91.7</td>
<td>81.8</td>
<td>61.98</td>
</tr>
<tr>
<td>SPPC</td>
<td>0.830</td>
<td>100</td>
<td>81.8</td>
<td>73.34</td>
</tr>
<tr>
<td>EWW</td>
<td>0.864</td>
<td>75.0</td>
<td>81.8</td>
<td>3.50</td>
</tr>
<tr>
<td>SWW</td>
<td>0.629</td>
<td>41.7</td>
<td>90.9</td>
<td>11.50</td>
</tr>
</tbody>
</table>

*Note.* AUC = area under curve. EPPC = English percent phonemes correct. SPPC = Spanish percent phonemes correct. EWW = English whole word. SWW = Spanish whole word.
Figure 2

*English Percent Phonemes Correct*
Figure 3

*Spanish Percent Phonemes Correct*
Figure 4

*English Whole Word*
A third aim of this study was to identify cut-off scores for each NWR task and scoring method that would yield optimal levels of sensitivity and specificity. Cut-off scores were determined from ROC analyses of each task as the points resulting in the highest sensitivity and specificity, with both being in the acceptable range. These scores are also reported in Table 4.
above. Cut-offs for PPC were 61.98 and 73.34 for English and Spanish, respectively. Whole word cut-offs were 3.50 and 11.50 for English and Spanish, respectively. Spanish cut-off scores were higher for both types of scoring.

**Classification Accuracy of Nonword Repetition Task Combinations**

We hypothesized that the English and Spanish NWR tasks used in combination would result in the best classification accuracy overall. To investigate this, binary logistic regressions were performed using the optimal cut-off scores identified above for each scoring method, yielding a predicted probability variable. ROC curves were created with this variable to calculate the AUC, as shown in Figures 6 and 7. Sensitivity, specificity, and total classification accuracy were calculated for combinations of English and Spanish PPC, and English and Spanish whole word scores using the optimal cut-off scores; see Table 5. For the combined language scores to identify a participant as DLD, the participant must have scored below the optimal cut-off for both tasks. The PPC combination yielded perfect specificity, but only 72.73% sensitivity. In contrast, the whole word combination resulted in acceptable levels for both sensitivity and specificity of 81.82% and 83.33%. Total classification accuracy was 86.96% for the PPC combination and 82.61% for the whole word combination.
Figure 6

Combined Percent Phonemes Correct

ROC Curve

Source of the Curve

Predicted probability

Sensitivity

1 - Specificity
Table 5

Diagnostic Accuracy of NWR Combinations

<table>
<thead>
<tr>
<th></th>
<th>AUC</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>Total Classification Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined PPC</td>
<td>0.924</td>
<td>72.73</td>
<td>100</td>
<td>86.96</td>
</tr>
<tr>
<td>Combined WW</td>
<td>0.864</td>
<td>81.82</td>
<td>83.33</td>
<td>82.61</td>
</tr>
</tbody>
</table>

*Note. AUC = area under curve. PPC = percent phonemes correct. WW = whole word.*
Discussion

The purpose of this study was to develop initial recommendations for the clinical use of NWR tasks by (a) determining the diagnostic accuracy of an English and Spanish NWR task; (b) determining the diagnostic accuracy of PPC and whole word scoring methods; and (c) identifying optimal cut-off scores for each task that resulted in the best sensitivity and specificity. This study also examined whether or not NWR tasks could be feasibly implemented in a clinical setting. The key findings and implications of the current study are discussed below.

Diagnostic Accuracy

The first finding is that the NWR tasks investigated in this study were diagnostically accurate. Previous studies pertaining to the diagnostic accuracy of NWR with DLLs have reported varying results. For instance, Windsor et al. (2010) found Dollaghan and Campbell’s (1998) English task to yield high sensitivity, but low specificity for Spanish-English DLLs. However, in the current study, both English and Spanish tasks were found to yield acceptable sensitivity and specificity using PPC scoring. Furthermore, Spanish PPC appeared to be the best measure for ruling in DLD, as it was the only measure to yield perfect sensitivity. This finding differed from previous studies, which found Spanish PPC to yield good specificity but poor sensitivity. For example, using the same Spanish task examined here, Windsor et al. (2010) reported a sensitivity of 58% and specificity of 82%. Gutiérrez-Clellen and Simon-Cereijido (2010) similarly found inadequate sensitivity (61%) but fair specificity for their Spanish NWR task. The authors of each of these studies concluded that Spanish NWR using PPC scoring could be used to rule out a disorder, but was insufficient to rule one in. Our results suggest greater potential for both Spanish and English PPC to be valid diagnostic tools for DLL children.
As a number of studies have reported promising results using whole word scoring, we anticipated that this scoring method would result in greater diagnostic accuracy than PPC. This work builds upon a recent study by Kelly (2021), who also compared NWR tasks in both Spanish and English using PPC and whole word scoring methods. Kelly (2021) found that Spanish whole word scoring yielded the best sensitivity and specificity of the four measures. Likewise, Guiberson and Rodríguez (2013) found whole word scoring to be more accurate than PPC. Contrary to these findings, neither task in the current study resulted in acceptable levels of both sensitivity and specificity using whole word scoring. This may be partially explained by the fact that Kelly (2021) investigated a different Spanish task than the one examined here. In that study, the mean whole word scores were 13.00% (2.08/16) for the TD group and 3.88% (0.62/16) for the DLD group. In contrast, the mean scores for this study were higher for both groups, at 46.65% (9.33/20) for the TD group and 40.90% (8.18/20) for the DLD group. This may indicate that the NWR list used here was not as difficult as the list used by Kelly. Additionally, while Guiberson and Rodríguez (2013) used the same nonword list, their participants were several years younger than the participants in this study. Given that performance on NWR has been shown to improve with age, it’s possible that an age effect contributed to differences in the diagnostic accuracy between age groups.

Because performance in both languages is typically a more accurate reflection of a child’s language ability, we hypothesized that a combination of English and Spanish tasks would result in the best diagnostic accuracy. This was the case for whole word scoring. While whole word scoring for the English and Spanish tasks individually resulted in acceptable specificity, sensitivity was not acceptable for either as a stand-alone measure. However, when the tasks were combined, whole word scoring resulted in both acceptable sensitivity and specificity, consistent
with findings from Kelly (2021) and providing additional support for the use of whole word scoring in the assessment of DLL children. Interestingly, the opposite effect was observed for PPC scoring when the English and Spanish tasks were combined. Specificity increased to 100%, but at a significant cost to sensitivity. While both English and Spanish PPC yielded good sensitivity individually, the tasks combined resulted in an unacceptable level. This occurred because participants were required to score below the optimal cut-off on both tasks to be classified as DLD by the combination PPC measure. This requirement resulted in perfect specificity; however, two participants who were misclassified as TD on only one task were consequently misclassified by the combination measure, thereby lowering the sensitivity. Had participants been required to score below the cut-off on only one task to be classified as DLD, sensitivity would have improved at the expense of specificity, as two TD participants would have been incorrectly classified as DLD. These results suggest that when using PPC scoring, a combination of tasks is effective for ruling out a language disorder, but it is less useful for ruling it in.

Another key outcome of this study is the identification of cut-off scores for each task that led to good diagnostic accuracy. While several studies have supported the accuracy of NWR as an assessment measure for language disorder in DLL populations, this is the first study to our knowledge to specify both PPC and whole word cut-off scores for Spanish-English DLLs. This is an important step towards NWR becoming a clinically useful measure for bilingual language assessment. The cut-off scores obtained here were higher for the Spanish task than the English task, congruent with observations from Windsor et al. (2010) that DLL participants consistently performed better on NWR in Spanish than in English, regardless of their classification as either DLD or typically developing. This pattern underscores the need for clinicians assessing bilingual
students to consider that DLLs tend to perform better in their native language in order to appropriately interpret assessment results. In addition, while Kelly (2021) noted that whole word scores may not be clinically practical for Spanish NWR due to the relatively low scores of both TD and DLD groups on the task, the optimal cut-off score for the task used in this study was much higher and therefore potentially more useful as a clinical measure.

Clinical Feasibility

A final question examined here was whether NWR tasks could be feasibly implemented as an assessment measure in the clinical setting. In the current study, tasks were administered in a naturalistic environment resembling a shared classroom or therapy room common in school settings. A primary concern pertaining to this testing location was whether noise or disruptions from other speakers would inhibit the ability to hear participant responses. However, of 109 audio recordings obtained from the larger sample, only seven (6.42%) were unable to be transcribed or scored due to poor audio quality or excessive background noise. Furthermore, NWR tasks were simple to administer and required no special equipment beyond a computer, phone, or other device with audio recording and playback capabilities. NWR was also efficient, as both the English and Spanish tasks could be administered together in fewer than four minutes. Whole word scoring, when used for a combination of English and Spanish tasks, could also be used with an acceptable level of classification accuracy while requiring less time to score than PPC. Each of these findings support the conclusion that NWR is a viable and efficient assessment measure that can be extended successfully to the clinical setting.

Recommendations for Clinical Use

The current study indicates that NWR tasks can be an appropriate assessment tool for practicing clinicians, particularly in school settings. While administration of NWR should occur
in a quiet location as much as possible to increase the accuracy of transcription and scoring, our findings suggest that it does not necessarily need to take place in a silent room as has typically been the case in prior research. Clinicians may additionally consider using a lapel microphone when recording child responses for later scoring to improve the overall quality of the recording.

The NWR tasks considered in this study were found to have good classification accuracy. While the greatest levels of sensitivity and specificity were identified using PPC scoring for English and Spanish tasks individually, clinicians are still advised to consider performance on both tasks, as best practice dictates assessment of both languages to truly represent a child’s language abilities (Levy et al., 2020; Peña et al., 2016). Sensitivity for the tasks combined using PPC scoring was less than desirable; however, specificity was perfect. Therefore, while PPC for both tasks is not sufficient to rule in a language disorder, clinicians can use this measure to rule out a disorder with a high degree of confidence.

Whole word scoring also shows potential to be clinically useful, but needs be explored further. The use of whole word scoring is not recommended when using English or Spanish tasks individually, due to the low accuracy obtained in this study. It does appear that whole word scoring can lead to good diagnostic accuracy when performance in both languages is considered, and given its quick scoring time, it may be a more appealing approach for practicing clinicians. However, additional research with larger sample sizes is needed to determine appropriate recommendations for this scoring method.

This preliminary analysis identified optimal cut-off scores for each NWR task that resulted in the best sensitivity and specificity. Based on these initial results, the cut-off scores recommended for the greatest accuracy using PPC scoring are 61.98% for English and 73.34% for Spanish. For whole word scoring, the recommended cut-off scores for English and Spanish
are 4 and 12, respectively. Please note that whole word cut-off scores have been rounded to the nearest whole number to reflect an actual score a child would receive. Future studies are needed to corroborate these scores in order to develop stronger recommendations.

Despite the promising ability of NWR to accurately discriminate between children with and without DLD, we caution against the use of NWR in isolation. As mentioned previously, clinicians are encouraged to use a converging evidence approach in their assessment of DLLs by synthesizing multiple sources of evidence (Castilla-Earls et al., 2020). It is our recommendation that NWR be included as one of several assessment tasks to obtain a clear picture of a child’s language ability.

**Limitations and Implications for Future Research**

There are several limitations to the current study which should be considered when interpreting and generalizing the results. The first is that this preliminary study consisted of a relatively small sample with only 23 participants. Second, although good effort was made to incorporate multiple sources of evidence in group determinations, the NLM did not effectively discriminate between DLD and TD groups. There was no significant difference in performance between groups, as all but one participant scored more than 1 SD below the mean. While the NLM has been shown to effectively discriminate language ability in bilingual populations (Petersen & Spencer, 2012), in this study it appeared somewhat biased against the DLL participants. One possible explanation for this is that the students in the normative sample completed two NLM tasks and the norms were derived using the higher of the two scores, which was usually obtained on the second administration (Petersen & Spencer, 2016). It is likely that a learning effect contributed to an improvement in scores between the first and second administration and, given that the participants in this study completed only one NLM task, could
explain why they performed poorly compared to the normative sample. Finally, while the testing environment in this study was designed to mimic a clinical rather than research setting, it is possible that ambient noise influenced performance on NWR tasks by reducing participants’ ability to hear and/or concentrate on the task. In addition, noise in the audio recordings could have limited the accuracy of the transcription and scoring of participant responses.

Future research of diagnostic accuracy and optimal cut-off scores should be conducted with larger sample sizes to either corroborate or clarify the results of this study. Future studies might also compare the diagnostic accuracy of multiple nonword lists for Spanish and English tasks to determine which tasks or combination of tasks are the most accurate and therefore most appropriate for clinical use. Since whole word scoring was performed in the current study following task administration rather than online, it would also be beneficial for future research to investigate the feasibility of whole word scoring in real-time, as this would further increase the overall efficiency and clinical utility of this scoring approach.

**Conclusion**

This study provided preliminary evidence to support the use of an English and Spanish NWR task as a language assessment measure for DLL children. Both tasks were determined to yield acceptable sensitivity and specificity using PPC scoring. Neither task resulted in adequate levels of both sensitivity and specificity as stand-alone measures using whole word scoring; however, whole word scoring did demonstrate acceptable diagnostic accuracy when performance in both languages was considered. Optimal cut-off scores were also identified for each NWR task and scoring method that resulted in the best sensitivity and specificity. The results of the current study suggest that NWR can be successfully implemented in the clinical setting as an
efficient, accurate measure of language ability. Further research is needed to develop stronger recommendations for its use in clinical practice.
References

https://www.asha.org/siteassets/surveys/demographic-profile-bilingual-spanish-service-members.pdf


https://doi.org/10.1044/2017_LSHSS-17-0013

https://doi.org/10.2167/beb392.0


https://doi.org/10.1017/S1366728914000248


https://doi.org/10.1044/1092-4388(2013/12-0056)


https://doi.org/10.1044/2019_JSLHR-L-18-0148


https://doi.org/10.1007/s12187-013-9230-6


APPENDIX A

Annotated Bibliography


Objectives: This study described the current practices and measures used by school-based speech-language pathologists (SLPs) in bilingual language assessment. It also discussed how closely the identified practices aligned with guidelines for best practice established by ASHA and the Individuals with Disabilities Education Act (IDEA).

Methods: School-based SLPs were recruited to complete an electronic survey regarding assessment practices and measures for bilingual children. Responses were obtained from 166 SLPs and were analyzed using descriptive statistics and content analysis.

Results: A majority of SLPs reported completing bilingual language assessment at least sometimes. Most respondents (73%) reported including a combination of formal and informal measures in their assessment. Sixty percent reported completing assessment measures in both languages often. Only 36% of respondents reported frequently analyzing language samples in both languages and 28% using dynamic assessment often. A lack of time was cited as the most common barrier to using dynamic assessment. The most commonly reported assessment tools included language sampling in English, standardized assessments in English, and standardized assessments in Spanish.

Conclusions: The current practices of school-based SLPs in bilingual language assessment have improved from past practices, though some areas of concern remain.
The results of this survey indicated that more SLPs recognize the importance of assessing the child’s performance in both English and their native language and report doing so more frequently. While more standardized assessments in languages other than English are being used more frequently, standardized English assessments remain the most commonly reported assessment tool. Additionally, the authors noted that many of the frequently used standardized assessments in Spanish report poor diagnostic accuracy or do not report it at all. They highlighted the need to exercise caution in interpreting the results of such assessments, as they may not provide an accurate representation of a child’s language ability.

*Relevance to the current study:* The bilingual assessment practices of school-based SLPs have improved in recent years to become more culturally competent; however, there is still a need for diagnostically accurate assessment tools that are quick and easy to administer. NWR is one such tool that has the potential to be a clinically useful measure, but cut-off scores and guidelines to extend its use into clinical practice must be established first.


*Objectives:* The purpose of this study was to investigate the diagnostic accuracy of nonword repetition (NWR), forward digit span (FWD), and sentence repetition (SRep) tasks to identify language disorder in Russian-Hebrew bilingual children using monolingual and bilingual cut-off scores.
Methods: Participants consisted of 230 school-aged children divided into six groups: Russian-speaking monolingual children with and without DLD, Hebrew-speaking monolingual children with and without DLD, and Russian-Hebrew bilingual children with and without DLD. All participants completed NWR, FWD, and SRep tasks in both Russian and Hebrew. A one-way ANOVA and post-hoc tests were used to examine differences in performance on each task between groups. Sensitivity and specificity for each task were determined at monolingual and bilingual cut-off points using a ROC analysis and likelihood ratios were then calculated.

Results: Significant differences between groups were observed for all tasks in both languages. NWR and SRep tasks in Russian and Hebrew were found to have adequate specificity and sensitivity for monolingual children. The monolingual cut-off scores for NWR did not yield acceptable levels of sensitivity and specificity in either language for the bilingual children, resulting in both over- and under-diagnosis. Bilingual cut-off scores improved the overall accuracy of NWR in both Russian and Hebrew. When performance on NWR in both languages was considered together using bilingual cut-off scores, overall accuracy improved to 89%, with excellent specificity (94%) but poor sensitivity (65%). A combination of NWR and SRep tasks in Hebrew using bilingual cut-off scores yielded 80% sensitivity and 93% specificity for bilingual children.

Conclusions: While all three tasks were concluded to have high accuracy for discriminating language disorder in monolingual children, monolingual cut-off scores were insufficient to accurately identify DLD in bilingual children. Bilingual cut-off scores improved overall accuracy, particularly when both languages were considered.
NWR in combination with other tasks using bilingual cut-off scores yielded adequate sensitivity and specificity for Russian-Hebrew bilingual children.

*Relevance to the current study:* Bilingual cut-off scores are necessary for NWR to be an accurate diagnostic tool for Spanish-English DLLs. The current study seeks to establish cut-off scores specific to this population in order to improve the clinical utility of NWR as one test measure for bilingual language assessment.


*Objectives:* The purpose of this study was to describe the performance of low-income DLL students on the CELF-4S and to determine the ability of the CELF-4S to accurately discriminate language disorder and typical language in this population.

*Methods:* Six hundred fifty-six low-income DLL children between ages 5;0 and 7;11 participated in the study. A subsample of 299 participants completed the Spanish Screener for Language Impairment in Children Morphosyntactic Task, the Structured Photographic Expressive Language Test – Third Edition (SPELT-3) in English, and Spanish language sample analyses to identify the presence of language disorder. Two hundred sixty-five children were found to have typically developing language, and 28 were identified as having a language disorder. The participants’ performance on the CELF-4S was compared to that of the normative sample. A ROC curve analysis was then used to determine the diagnostic accuracy of the CELF-4S for low-income DLL children.
**Results:** The sample’s average core language score on the CELF-4S was more than 1 SD below the normative sample mean. A majority of the children in the study sample scored below 1 SD below the mean and would therefore be misidentified as having DLD using the CELF-4S normative values. The CELF-4S was estimated to have 93% sensitivity and 65% specificity using 85 as the cut-off score (as recommended by the manual). However, acceptable levels of sensitivity and specificity were achieved using a cut-off score of 78.

**Conclusions:** The CELF-4S frequently over-identified low-income DLLs as having language disorder. Despite its frequent use in clinical practice, clinicians should use caution when interpreting scores on the CELF-4S, particularly for students with low SES backgrounds. Converging evidence should be used rather than a single assessment measure to make appropriate diagnostic decisions.

**Relevance to the current study:** NWR may be a less biased measure of language ability for DLL children and may be less influenced by risk factors such as SES than knowledge-dependent measures. NWR could make an important contribution as one source of evidence in bilingual language assessment.


https://doi.org/10.2167/beb392.0

**Objectives:** This article reviewed differences in language acquisition between monolingual and bilingual children and discussed the inability of traditional language assessments to accurately identify language impairment in bilingual populations.
Conclusions: Bilingual children are at greater risk of misidentification for language disorder and are simultaneously over- and under-identified in the U.S. public schooling system. Language acquisition patterns differ across languages, and language acquisition for bilingual children is often dependent on the child’s exposure to each language as well as cross-linguistic effects. While typically developing bilingual children generally know the same number of words as same-age monolingual peers, their vocabulary is often distributed across both languages. Thus, their word knowledge in either language can appear smaller than monolingual peers. Mixed knowledge of the morphosyntactic rules of each language can also result in unusual grammatical productions in either language. Bilingual children demonstrate similar error patterns as monolingual children with language disorder, as well as unique error patterns.

Traditional language assessments, including developmental milestones and standardized tests, are inadequate measures for the identification of language disorder in bilingual children. Bilingual children follow a different pattern of language acquisition than their monolingual peers, and developmental milestones for acquisition of their first or second language are therefore inappropriate standards of comparison for their language performance. Standardized tests are typically normed on monolingual populations, and likewise do not reflect typical development for bilingual children. Furthermore, many commonly used standardized language assessments do not have adequate classification accuracy for identification of language disorder.

The authors recommended development of additional assessment tools that accurately identify clinical markers for language disorders in bilingual populations in order to prevent misdiagnosis of bilingual children.
Relevance to the current study: The current study seeks to develop clinically useful recommendations for the inclusion of NWR tasks in the language assessment of bilingual children. NWR would provide a culturally appropriate assessment for bilingual children that primarily examines language processing abilities rather than language knowledge to form appropriate diagnostic decisions.


Objectives: The purpose of this study was to investigate the diagnostic accuracy of various single assessment measures and combinations of measures for identifying language disorder in bilingual children learning Italian as a second language.

Methods: Fifty-five bilingual children (35 TD; 20 DLD) with at least 2 years of exposure to Italian participated in this study. The children completed an Italian NWR task as well as formal language assessments in Italian to obtain measures of their expressive and receptive morphosyntax and grammar, receptive vocabulary, and narrative skills. To collect information regarding the children’s linguistic competence in their native language, the children completed the Prove BaBIL test and their parents completed the Alberta Language and Development Questionnaire (ALDeQ). Discriminant analyses were performed to obtain sensitivity and specificity values for single measures and combinations of measures.
Results: Significant group differences were observed for measures of both L1 and L2, with the DLD group performing more poorly compared to the TD group. No single measure of L2 performance yielded adequate levels of sensitivity and specificity. L2 morphosyntactic comprehension achieved an overall accuracy of 80% but low specificity (77.10%). NWR yielded high specificity but low sensitivity, with overall accuracy of 76.4%. The ALDeQ achieved adequate overall accuracy (89.1%), with good sensitivity and fair specificity. The greatest level of accuracy and efficiency was achieved using a combination of measures from the ALDeQ, L1 vocabulary, L2 morphosyntactic comprehension and production, and L2 NWR. This combination yielded 100% sensitivity and 91.4% specificity. Discriminant accuracy was found to be low (78.2%) when measures of L1 were not considered.

Conclusions: Most single measures of language ability had poor discriminant accuracy. The ALDeQ and L2 morphosyntactic comprehension reached fair levels of accuracy independently. Overall accuracy improved significantly when multiple measures were considered. The parent questionnaire for L1, NWR in L2, and expressive and receptive morphosyntax in L2 appeared to be the most informative measures. The results indicated that measures of language ability in both L1 and L2 are necessary to achieve good diagnostic accuracy.

Relevance to the current study: Greater levels of diagnostic accuracy for DLL children are achieved when multiple measures are examined and performance in both languages is considered. The current study seeks to expand the informative value of NWR as a measure for bilingual language assessment by considering multiple NWR tasks in both L1 and L2.

**Objectives:** The purpose of this study was to compare the performance of minority and majority children on knowledge-dependent and processing-dependent language measures.

**Methods:** Forty-nine majority and 107 minority children between the ages of 11 and 14 completed one knowledge-dependent task (the Oral Language Scale from the Woodcock Language Proficiency Battery-Revised) and three processing-dependent language tasks (NWR, the Competing Language Processing Task, and the shortened version of the Revised Token Test). Analyses of variance (ANOVA) were used to examine differences in performance between the two groups for each task.

**Results:** No statistically significant differences were observed between groups for any of the processing-dependent tasks. However, the minority group performed significantly below the majority group on the knowledge-dependent task.

**Conclusions:** The differences in performance between groups on the knowledge-dependent task but not on any of the processing-dependent tasks suggest that the knowledge-dependent task was biased against the minority children. Processing-dependent tasks may be more appropriate measures of language ability than knowledge-dependent tasks in order to minimize bias against children with different backgrounds and experience.

**Relevance to the current study:** NWR is one processing-dependent measure that has the potential to reduce assessment bias against Spanish DLL children. More
information regarding its use and discriminant accuracy is needed before NWR can be applied to clinical practice.


*Objectives:* This article reviewed four language assessment measures and provided a tutorial for using converging evidence to make appropriate diagnostic decisions for dual language learners.

*Conclusions:* A converging evidence approach synthesizes data from multiple assessment measures such as language experience questionnaires, language sample analyses, dynamic assessment or response to intervention, and standardized tests, and is the most culturally appropriate assessment approach for dual language learners. Evaluating patterns of performance across multiple sources of evidence provides a greater degree of diagnostic accuracy and confidence for identifying language disorder in dual language learners than does relying on a single assessment measure.

*Relevance to the current study:* The current study aims to determine the accuracy of nonword repetition tasks to identify dual language learners with language disorder. A NWR task with high diagnostic accuracy would provide useful data that, in combination with other language measures, could increase confidence in diagnostic decisions for Spanish-English bilingual children.

**Objectives:** This systematic review evaluated and compared the psychometric quality of standardized language assessments for school-aged children.

**Methods:** A total of 15 assessments were examined in this review. Psychometric information about each assessment was obtained from assessment manuals and 121 independent studies of various psychometric properties of the assessments. The Consensus Based Standards for the Selection of Health Status Measurement Instruments (COSMIN) four-point checklist was used to rate nine psychometric properties for each study.

**Results:** Of the 121 studies included in this review, only 60 were rated as having *good* or *excellent* methodological quality. Approximately half of the included studies had significant methodological flaws that limited their informative value. Additionally, some of the findings from independent studies conflicted with information reported in the assessment manuals. The Assessment of Literacy and Language, the Clinical Evaluation of Language Fundamentals-5th Edition, the Clinical Evaluation of Language Fundamentals-Preschool: 2nd Edition, and the Preschool Language Scales-5th Edition were determined to have better psychometric evidence than the other assessments included in this study; however, none of the assessments could be described as having *good* psychometric evidence from the information available.

**Conclusions:** Currently, the available research suggests that the 15 assessments evaluated in this study are limited in their psychometric quality. Based on the available
evidence, the authors recommended that the ALL, CELF-5, CELF-P:2, and PLS-5 be used for the language assessment of school-aged children. Further research of psychometric quality of language assessments is needed for the appropriate use of standardized assessments in clinical practice.

Relevance to the current study: Many available standardized language assessments lack good psychometric properties and are therefore limited as a clinically informative measure. Other sources of evidence, such as NWR, are needed to draw accurate conclusions about a child’s language ability and make appropriate diagnostic decisions.


**Objectives:** This study investigated the ability of a nonword repetition (NWR) task specifically designed to be independent of language experience to differentiate between children with and without language disorder. It also compared the diagnostic accuracy of the NWR task with that of a language-dependent, norm-referenced assessment to identify children with language disorder.

**Methods:** A total of 16 nonwords were constructed to avoid English word-likeness, consisting of four nonwords each for one, two, three, and four syllables. All nonwords contained initial and final consonants and avoided consonant clusters and late-developing phonemes. Only tense vowels were included in the nonwords, in order to avoid reducing vowels to a schwa and therefore confidently attribute errors to incorrect recall rather than incorrect perception. The nonwords did not conform to typical stress
patterns or include consonants in positions where they occurred more than 25% of the
time in English words. Additionally, no phonemes were repeated in the same nonword.
The nonwords were presented to a group of 40 children between the ages of 6;0 and 9;9,
including 20 children diagnosed with language disorder by a certified SLP and 20
typically developing children. The nonwords were presented from a recording at a
consistent rate, in order of increasing syllable length. The children were instructed to
repeat each nonword after its presentation, and their responses were recorded and
transcribed. They were then scored by accuracy for each phoneme. Incorrect scores were
given for phoneme substitutions and omissions, but not for distortions or additions. Each
child’s percentage of phonemes correct (PPC) was then calculated for each nonword
length and for the nonwords overall. ANOVA and Scheffe tests were used to examine
and compare each group’s performance on the task overall and for each syllable length.

The second part of this study included 85 children ages 5;8 to 12;2 and consisted
of 41 typically developing children and 44 with language disorder, determined by
enrollment in language intervention. These children participated in the same NWR task
described above, and were also administered the Test of Language Development-2
(TOLD-2). The children’s PPC for all nonword lengths and in total were calculated, as
well as their Spoken Language Quotients (SLQ) from the TOLD-2. Likelihood ratios for
PPC and the TOLD-2 were calculated and compared.

Results: The authors found that children with language disorder in the first part of
this study obtained significantly lower PPCs for three- and four-syllable nonwords and
for the nonwords in total compared to the children without language disorder.
Furthermore, no overlap of performance was found between the groups for three- and
four-syllables and total PPC using 99% confidence intervals. These differences in performance could not be attributed to limited phonetic inventories or to differences in non-verbal IQ. In the second part of this study, the authors determined that the total PPC was the most accurate measure from the NWR task to identify children with language disorder and distinguished between the two groups with near perfect accuracy. PPC for three- and four-syllables likewise had high diagnostic accuracy. In contrast, SLQ from the TOLD-2 yielded lower likelihood ratios for identifying children with language disorder, and weak likelihood ratios for identifying children without language disorder.

Conclusions: The results of this study indicated that children with and without language disorder perform differently on the NWR task investigated here. Furthermore, the NWR task was more accurate in distinguishing between children with language disorder and typically developing children than the TOLD-2. The authors concluded that the diagnostic accuracy of the NWR task, in addition to its ease and brief time to administer, made it much more useful as a screening measure than the TOLD-2.

Relevance to the current study: This study provided preliminary evidence for the diagnostic accuracy of NWR and its clinical utility for identifying children with language disorders. These nonwords are included in the current study.


Objectives: This article provided a meta-analysis of various language assessment measures and their accuracy for identifying language disorder in Spanish-English bilingual children.
Methods: The authors searched two electronic databases for articles concerning bilingual language assessment. Nine studies which contained sufficient information to calculate metrics of diagnostic accuracy and confidence intervals for a total of 17 language measures were included in the meta-analysis.

Results: All but one language measure examined in the analysis yielded positive likelihood ratios in the suggestive range (>3.0). However, the confidence intervals for likelihood ratios for all measures were broad and often approached or extended into the uninformative range.

Conclusions: No single measure was identified as the most accurate indicator of language disorder or typical language in Spanish-English bilingual children. Individual measures could be considered to be suggestive of a language disorder, but would require additional clinical information to support a diagnostic decision. Further studies of the diagnostic accuracy of various language measures are also needed.

Relevance to the current study: Multiple language measures are necessary in order to make valid diagnostic decisions in bilingual language assessment. NWR may be one informative measure for this purpose. The current study seeks to add to existing research of the diagnostic accuracy of NWR for Spanish-English bilingual children.


Objectives: The purpose of this study was to examine which assessment tools school-based SLPs use for the assessment of DLLs and to determine whether SLPs primarily
rely on single measures or a combination of measures in their assessment of individual children.

Methods: School-based SLPs were recruited via email and at a state convention to complete a survey regarding assessment practices in schools. The responses from 222 SLPs who met the inclusion criteria (i.e., had past or present experience with DLL children) were then analyzed using frequency and correlational analyses.

Results: Case history and observation were the most frequently reported assessment tools used in the language assessment of DLLs. Standardized assessments were used occasionally, often, or always by 86% of respondents. Language samples were collected often or always by a majority of SLPs. Less frequently reported assessment tools included dynamic assessment and criterion-referenced measures. All but one respondent (99.5%) reported using multiple assessment tools for each child they evaluated.

Conclusions: The results of this survey indicated that most school-based SLPs follow guidelines to include multiple assessment tools in their evaluation of DLL children. However, the authors noted that only respondents with experience with DLLs were included in the study. Respondents who were excluded from the study due to a lack of experience with this population might report different assessment practices.

Relevance to the current study: School-based SLPs appear to be using multiple assessment tools frequently in their assessment of DLL children. More unbiased and accurate assessment tools, such as NWR, should be included in this practice to further inform diagnostic decisions.

**Objectives:** The purpose of this study was to develop a set of Spanish nonword stimuli to be used for nonword repetition (NWR) tasks and to examine patterns of performance between Spanish-English bilingual children on the task.

**Methods:** A set of 20 Spanish nonwords was developed using similar standards as those from Dollaghan and Campbell (1998) and were constructed to avoid lexical influence as much as possible. Four nonwords were constructed for each length between one and five syllables. The nonwords conformed to typical Spanish stress and syllable patterns and contained 11 consonants and five vowels used in Spanish. Consonant clusters and later-developing consonants were excluded. The frequency of consonants included in the nonwords corresponded to the frequency of occurrence in the language. Participants of this study included fourteen Spanish-English bilingual preschool children who were divided into two groups according to their age: 3;6-4;0, and 4;3-5;6. The children were administered the PLS-4 in Spanish and English and participated in the NWR task. The children were asked to repeat each nonword exactly and their responses were transcribed and scored by a native Spanish speaker. Each phoneme was scored as either correct or incorrect. Omissions and substitutions were counted as incorrect, while additions and distortions were not. The total PPC and PPC for each syllable length was calculated for each child. A two-way ANOVA was conducted to investigate the effects of age and syllable length on repetition accuracy. Correlations between NWR performance and PLS-4 scores were also calculated.
Results: Significant effects for both age and length were found on repetition accuracy. The older group of children yielded a higher accuracy than the younger group for four- and five-syllable nonwords. The younger group’s accuracy decreased significantly for nonwords with three or more syllables. The authors found a significant partial correlation between Spanish and English scores on the PLS-4, but did not find significant correlations between NWR scores and PLS-4 scores in either Spanish or English.

Conclusions: The nonwords constructed for this study were concluded to be an appropriate difficulty and repetition performance corresponded with developmental expectations. Performance decreased with increasing nonword length, and older children performed better than younger children. Additionally, proficiency in either language, as measured by the PLS-4, was not significantly correlated to performance on the NWR task, suggesting that the nonwords tapped more language processing skills rather than specific language knowledge. The authors concluded that this particular set of Spanish nonwords can be used for NWR tasks with young children as a measure of language processing abilities.

Relevance to the current study: The Spanish nonwords developed for this study are an appropriate measure of language processing rather than language knowledge for typically developing bilingual children, and are included in the current study.


Objectives: This article described a theoretical framework outlining the relationship between nonword repetition and language learning.
Conclusions: Multiple studies have reported strong relationships between vocabulary knowledge and the ability to repeat nonwords accurately, particularly during the early stages of language acquisition. Children with language disorders perform poorly on NWR compared to typically developing children; thus, poor NWR accuracy is often considered a clinical hallmark of DLD. NWR taps multiple processes including auditory processing, phonological processing, phonological storage, speech-motor planning, and speech output. Phonological storage, which is presumed to be an area of deficit for children with DLD, is highly involved in the acquisition of novel phonological forms. Deficits in phonological storage therefore result in slower word-learning. Such a deficit can also impair NWR ability as the demands of the task exceed the resources available, which is why children with DLD demonstrate significant difficulty repeating nonwords of increased length. In addition, impairments in verbal working memory and increased processing demands likely contribute to an NWR deficit. NWR ability appears to have a genetic basis and is not influenced by external factors such as maternal education. Thus, NWR is considered to be a less biased measure of language ability.

Relevance to the current study: This article explains the theoretical basis for NWR as a potential assessment tool to identify children with DLD.

Objectives: This study investigated the respective contributions of phonological structure and language experience to bilingual children’s performance on NWR, and examined the influence of the task language on patterns of accuracy for each syllable length.

Methods: Fifty-two typically developing Spanish-English bilingual kindergarteners participated in this study. The children were matched on first exposure to English and inverse input of each language at the time of the study, so that the children had same lengths of exposure to English but differed by current levels of use. The children participated in Spanish and English NWR tasks and their accuracy was scored by calculating their percentage of phonemes correct for each syllable length.

Results: No significant difference in performance was found between the Spanish and English tasks for one, two, and three-syllable nonwords. However, performance on four-syllable nonwords was significantly better on the Spanish task than in English. The Spanish-dominant group had higher PPC scores in both languages than the English-dominant group. No statistically significant interactions were found between language dominance and syllable length, or between the task language and language dominance.

Conclusions: This study indicated that phonological structure and language experience influence performance on NWR tasks. The results suggest the presence of a frequency or practice effect such that more practice or exposure to multisyllabic words in Spanish transferred to better performance on longer nonwords in both languages.

Relevance to the current study: Spanish-English bilingual children are likely to perform better on NWR tasks in Spanish than in English. Typically developing children with significant Spanish exposure also tend to perform better on longer nonwords than
children with less Spanish exposure. These patterns of performance are expected to hold in the current study.


**Objectives:** The purpose of this study was to specify cut-off scores and determine the diagnostic accuracy of English assessment measures for identification of bilingual children with language disorders.

**Methods:** Participants consisted of 167 Spanish-English bilingual children in the first grade. Twenty-one students were identified as having a language disorder by three bilingual SLPs using a severity rating scale for each language. The children completed subtests of the Test of Language Development – Primary: 3rd Edition (TOLD-P:3) and the Test of Narrative Language (TNL). Following the EpiSLI model, the standard scores were then used to produce five language composite scores (semantics, syntax, narration, comprehension, and production). ROC analyses and logistic regression were used to determine the sensitivity and specificity of the original EpiSLI diagnostic criteria and to identify new cut-off scores to yield optimal diagnostic accuracy.

**Results:** The original EpiSLI diagnostic criteria (standard score of 6.25 or below on two or more composite scores) yielded high sensitivity (.95) but poor specificity (.45) and were therefore uninformative for identifying language disorder in the bilingual children. Using the same model with revised cut-off scores improved specificity to .68,
but reduced sensitivity to .86. No single subtest yielded adequate sensitivity and specificity.

Conclusions: Bilingual children were more likely to be misdiagnosed with a language disorder when using the monolingual diagnostic criteria than the revised cut-off scores. Increased accuracy of English testing may be achieved for bilingual children when alternative cut-off scores are applied.

Relevance to the current study: Like alternative cut-off scores, NWR is another method to reduce assessment bias and inform the diagnosis of language disorder in bilingual children.


Objectives: The purpose of this study was to identify and compare rates of misdiagnosis of language disorder between monolingual and bilingual children.

Methods: German-speaking children and children learning German as a second language were included in this study. The presence of four risk factors and past clinical diagnoses of language disorder by an SLP were noted in parent questionnaires and interviews. The participants were then identified as typically developing or having a language disorder based on their performance on a standardized language assessment in German (LiSe-DaZ). This was compared to their clinical diagnosis using chi-square tests to calculate rates of misdiagnosis.

Results: Both the monolingual and bilingual groups experienced a similar prevalence of risk factors of language disorder. Misdiagnosis was common among both
groups, with underdiagnosis occurring more frequently for both monolingual and bilingual children. Overdiagnosis was more common among the bilingual group. The rate of overdiagnosis for bilingual children was 27.3%, while the rate of overdiagnosis for monolingual children was 14.5%.

**Conclusions:** While children learning German as a second language were not more likely to present with a language disorder, they were more likely to be over-diagnosed compared to their monolingual peers.

**Relevance to the current study:** Spanish-English bilingual children are similarly at risk of over- and underdiagnosis; therefore, more accurate assessment tools are needed to reduce the rate of misdiagnosis for bilingual children.


**Objectives:** This study examined the training, practices, and challenges of SLPs serving culturally and linguistically diverse (CLD) children.

**Methods:** Practicing SLPs in Colorado were recruited to complete an electronic survey consisting of questions pertaining to background information, diversity training, and professional perspectives. Descriptive statistics were calculated to summarize the responses for each survey question.

**Results:** Ninety-seven percent of respondents were white, and only 11% were bilingual. Approximately half of SLPs reported a large number of Spanish-speaking children on their caseloads. Most respondents felt comfortable working with culturally and racially diverse students, but only 51% felt competent in their assessment and
treatment of bilingual students. Fewer SLPs reported using standardized assessments in English for CLD children than in previous surveys. However, 75% of respondents indicated that a major challenge to bilingual assessment was a lack of appropriate assessment tools for this population.

Conclusions: While SLPs have generally improved in their provision of services to CLD children, many still do not feel comfortable in their assessment and treatment of bilingual children. This may be largely due to the lack of appropriate assessment tools for CLD children.

Relevance to the current study: The current study seeks to develop the clinical utility of NWR as an additional assessment tool appropriate for CLD children.


Objectives: The purpose of this study was to examine and compare performance on a Spanish NWR task for Spanish-speaking preschool children, and to compare the classification accuracy of the task using different scoring methods.

Methods: Forty-five preschool-age Spanish-speaking children with and without language impairment participated in a Spanish NWR task. The nonwords were presented live rather than via a recording in order to engage the young children. The children were asked to repeat what a puppet said, and the accuracy of their responses was scored by calculating their percentage of phonemes correct and item-level accuracy.

Results: Item-level scoring was found to have a greater classification accuracy than PPC scoring in identifying children with language disorder. PPC scoring led to poor
sensitivity and adequate specificity, while item-level scoring yielded adequate sensitivity and specificity. The greatest differences in performance between the typically developing and language impaired groups were observed for five-syllable nonwords.

Conclusions: NWR tasks can be successfully implemented with preschool children and through different modes of administration. In order to be a clinical useful measure, NWR tasks should include nonwords consisting of three to five syllables. NWR can contribute to identification of language impairment in Spanish-speaking children when item-level scoring is employed.

Relevance to the current study: The results of this study support the use of NWR tasks with young, Spanish-speaking children. Additionally, the findings indicate that item-level scoring is a more accurate measure to identify language disorder than PPC scoring. The current study will further investigate the accuracy of both scoring methods in discriminating children with language disorder from typically developing children.


Objectives: This study investigated the ability of Spanish and English nonword repetition tasks to accurately differentiate between bilingual children without language impairment. The influence of language skill on classification accuracy of each measure was also examined.
**Methods:** One hundred forty-four bilingual children with various levels of English proficiency and language dominance participated in this study. Ninety-five children were determined to have typically developing language and 49 were determined to have language impairment based on parent/clinical concern and performance on morphosyntax measures. The children participated in NWR tasks in each language on different days. Their responses were analyzed to calculate the number of errors and total percentage of phonemes correct (PPC). Differences in PPC between groups were compared, and ROC curves were used to determine which cut-off scores provided the greatest sensitivity and specificity for each task. Likelihood ratios were then calculated for those cut-off scores.

**Results:** Children with language impairment scored significantly lower on the NWR tasks in both English and Spanish. Some children performed better on the task in their non-dominant language. Additionally, many children with typical language scored below the cutoff score in one language, but not the other. While specificity was moderate for both tasks at the determined cut-off scores, sensitivity was poor. However, specificity and positive likelihood ratios were high when performance on both tasks was examined together.

**Conclusions:** The results of this study indicated that NWR tasks in only one language are not clinically useful to differentiate bilingual children with and without language disorder. Bilingual screening and assessment measures are supported by these findings in order to account for differences in language skill across languages. NWR tasks in both languages, in conjunction with other language measures, could increase identification of bilingual children with language disorders.
Relevance to the current study: NWR tasks have high diagnostic accuracy for identifying bilingual children with language impairment when performance on the task in both languages is considered. NWR in only one language, regardless of the child’s language dominance, is not sufficient to differentiate bilingual children with and without language disorder.


**Objectives:** The purpose of this study was to compare the diagnostic accuracy of an English and Spanish NWR task using two different scoring methods. This study also compared error patterns between children with and without DLD.

**Methods:** Twenty-six Spanish-English bilingual children with and without DLD completed a NWR task in both Spanish and English. Percent phonemes correct (PPC) and whole word scoring were both employed and a discriminant function analysis was performed to calculate the sensitivity and specificity for each scoring method. Repeated measure ANOVAs were used to compare error patterns between groups.

**Results:** The Spanish NWR task predicted group placement more accurately than the English task. Whole word scoring also resulted in greater diagnostic accuracy than PPC; however, participants in both groups repeated few whole words correctly, potentially limiting the clinical utility of this scoring method. Children with DLD made substitution and omission errors more frequently than children with typically developing language.
Conclusions: Spanish NWR can be an effective measure to identify bilingual children with language disorder. Whole word scoring may improve classification accuracy but may not be clinically practical.

Relevance to the current study: The current study builds upon this previous work by further comparing the diagnostic accuracy of multiple NWR tasks using both scoring methods.


Objectives: This study examined the comfort levels and professional development experiences of SLPs working with ELL students.

Methods: SLPs were randomly selected from the ASHA Schools Conference and the Missouri Speech-Language-Hearing Association to complete a brief survey regarding their comfort levels in assessing and treating ELL students as well as their professional development experiences. One hundred ninety-two SLPs completed the questionnaire and responses were then analyzed using descriptive statistics.

Results: Most SLPs who participated in the study indicated that they felt uncomfortable assessing ELL students, regardless of their years of experience. Significant correlations were found between SLPs’ comfort levels in assessing and treating ELL students and the number of professional development workshops attended.

Conclusions: Many SLPs do not feel comfortable providing services to ELL children. Additional training in multicultural service provision and second language
acquisition is recommended to increase SLPs’ comfort levels when working with this population.

*Relevance to the current study:* Increasing the availability of culturally sensitive assessment tools is likely to increase SLPs’ comfort levels in assessing ELL children. The current study will help develop clinical guidelines for NWR to aid SLPs in their assessment of Spanish-English bilingual children.


*Objectives:* The purpose of this study was to examine and compare the accuracy of a knowledge-dependent task and a nonword repetition task in discriminating monolingual children with language impairment from typically developing bilingual children.

*Methods:* This study included 100 children between ages 7 to 13. The children composed three different groups: monolingual English speakers with language impairment, typically developing monolinguals, and typically developing Spanish-English bilinguals. All children participated in a NWR task using nonwords developed by Dollaghan and Campbell (1998) and were scored by their percentage of phonemes correct for each syllable length. The children were also administered the Competing Language Processing Task (CLPT), during which they were directed to judge the truth of three-word declarative statements immediately after its presentation, and then recall the final word of each sentence following every set of two statements. The children’s percentage of correct responses for comprehension and word recall were then calculated. The
influence of age on performance for each task and between-group differences of performance were analyzed and likelihood ratios for each measure were calculated.

Results: The analysis found no statistically significant differences in comprehension on the CLPT between the typically developing and language impaired groups. The typically developing monolingual group performed significantly better on the recall task. Differences between the bilingual and language impaired groups on the recall task were not statistically significant. On the NWR task, the typically developing monolingual group performed significantly better than the other two groups, and the bilingual group performed significantly better than the language impaired group. The greatest differences in performance between groups occurred for four-syllable nonwords. While the likelihood ratios were calculated at cut-off scores to obtain maximum sensitivity and specificity, likelihood ratios for the CLPT recall task did not have adequate accuracy in discriminating language impairment. Likelihood ratios for the NWR task indicated greater accuracy in discriminating between typically developing and language impaired monolingual children than for identifying all three groups.

Conclusions: The findings of this study indicated that the CLPT recall task did not accurately discriminate the bilingual group from the other two groups and was not a clinically useful measure. Significant differences in performance on the NWR task were found between the bilingual and typically developing monolingual groups, suggesting that the specific nonwords used in this study were somewhat biased against CLD children, despite it being a processing-dependent task. The authors concluded that NWR can rule out language impairment but cannot always rule it in for CLD children.
Additionally, for NWR to be diagnostically useful, it should focus on nonwords of four or more syllables, as all groups demonstrated significant overlap for shorter lengths.

Relevance to the current study: Typically developing bilingual and monolingual children perform differently on NWR tasks, supporting the need for NWR in both languages to accurately discriminate between bilingual children with and without language disorder. NWR tasks should include nonwords with four or more syllables in order to achieve greater diagnostic accuracy.


Objectives: This study examined the assessment practices of SLPs for Spanish-English dual language learners, their adherence to guidelines, and the occurrence of misdiagnosis.

Methods: A records abstraction review was conducted of 88 reports documenting a language disorder. Reports were reviewed for specific assessments, assessment language, the child’s language proficiency, home language, English exposure, and schoolwork.

Results: SLPs most commonly used standardized assessments in English and Spanish. None reported the use of non-standardized assessment measures or included information regarding the child’s language dominance, home language, English exposure, or quality of classwork. Only one involved an interpreter in the assessment.

Conclusions: Common assessment practices of SLPs are likely to misidentify Spanish-English DLLs and often do not adhere to guidelines to test children in both languages.
Relevance to the current study: NWR is a potential assessment measure that is more appropriate for CLD students and could help reduce the risk of misdiagnosis.


**Objectives:** The purpose of this study was to describe the perspectives and beliefs of SLPs across the United States regarding bilingual language assessment.

**Methods:** Monolingual and bilingual SLPs from five states were recruited to complete a questionnaire consisting of 25 questions regarding their own language experience, training, decision-making, and beliefs about bilingual assessment. Respondents were divided according to their responses to compose three groups: monolingual SLPs (M SLPs), those who learned a second language in an academic setting (AS SLPs), and those who learned a second language via cultural experience, such as at home or abroad (CE SLPs). ANOVAs were used to analyze differences in responses between the three groups. Qualitative analyses were also used to summarize comments written by the respondents.

**Results:** CE SLPs reported the greatest personal efficacy in bilingual assessment, while the M and AS groups did not significantly differ in their personal efficacy. However, the majority of respondents in all groups reported low personal efficacy (*not competent* or only *somewhat competent*). Additionally, 40% of respondents indicated they would be less likely to recommend language intervention for a bilingual student than they would for a monolingual student.
**Conclusions:** The majority of SLPs do not feel competent in their language assessment of bilingual children. Many SLPs are also likely to be more conservative in their diagnostic decisions and treatment recommendations for bilingual children than for monolingual children. This may result in underidentification of bilingual children with language disorders who would benefit from intervention.

**Relevance to the current study:** Many SLPs report low competency in their bilingual assessment practices and decision-making. NWR would be a helpful tool to guide clinicians in their assessment and recommendations for bilingual children.


**Objectives:** The purpose of this study was to determine and compare the classification accuracy of a standardized assessment developed for Spanish-English dual language learners (SEDLLs) with that of translated assessments in either language. It also investigated whether or not classification accuracy would increase with the addition of informal measures.

**Methods:** Thirty SEDLLS with and without language disorder were recruited from a public school to participate in this study. The children completed the morphosyntax and semantics subtests of the Bilingual English-Spanish Assessment (BESA) and the Spanish and English versions of the Clinical Evaluation of Language Fundamentals Preschool-Second Edition (CELF-P2). They also participated in a narrative retell and a dynamic assessment task. ANOVAs were used to examine differences
between groups for each task and discriminant analyses were performed to calculate the classification accuracy for each measure and combination of measures.

**Results:** The BESA was found to have good sensitivity and fair specificity, with 90% classification accuracy overall. The CELF-P2 in Spanish yielded the same overall accuracy but had lower sensitivity and higher specificity. The greatest overall accuracy was achieved using a combination of the BESA and MLU from the narrative retell. This combination yielded 100% sensitivity and 92.9% specificity.

**Conclusions:** The BESA as a single measure had adequate classification accuracy to identify bilingual children with and without language disorder. Combining this measure with MLU from a language sample resulted in excellent sensitivity and specificity.

**Relevance to the current study:** The BESA is an appropriate assessment tool to discriminate between language disorder and typical language in Spanish-English bilingual children.


https://doi.org/10.1044/2020_PERSP-19-00115

**Objectives:** This review article discussed the implications of misidentification of bilingual and multilingual children with language disorder and examined differences between languages in multiple domains. It described evidence-based approaches to assessment that can reduce the risk of misdiagnosis in this population.
Conclusions: Second-language learners are at risk of both over- and underidentification as having a language disorder. These children are sometimes overidentified because their language performance in their second language is still developing and may resemble the performance of a monolingual child with a language disorder. Underidentification can result from a conscious effort of the SLP to avoid overidentifying culturally and linguistically diverse children as having a language disorder. Consequences of underidentification of children with a language disorder include later reading and academic challenges. Clinicians must be culturally competent and aware of the differences between languages in order to accurately distinguish between difference and disorder and avoid misdiagnosis. Evidence-based approaches to bilingual language assessment include assessment of both languages, conceptual vocabulary scoring, contrastive analysis of speech-language errors, norm-referenced assessments developed for bilingual populations, parent-teacher questionnaires, dynamic assessment, and nonword repetition.

Relevance to the current study: The aim of the current study is to establish guidelines and cut-off scores for a nonword repetition task to be used in assessment of Spanish-English bilingual children.

https://doi.org/10.1044/jshd.4901.34

Objectives: The purpose of this review was to examine the psychometric quality of standardized language and articulation assessments used with preschool children.
Methods: A total of 30 language and articulation assessments were included in this review. Using the test manuals and other information from the test, each assessment was evaluated according to ten psychometric criteria.

Results: Only three of the 30 assessments met five or more criteria. Half of all assessments included in the review met two criteria or fewer.

Conclusions: The majority of these standardized assessments have poor psychometric quality. Many fail to report important information to determine if certain criteria are met. Clinicians must consider the psychometric flaws of the tests they use and interpret test results with caution.

Relevance to the current study: Many standardized language assessments are insufficient to confidently confirm or rule out the presence of a language disorder due to psychometric weaknesses in their design. Thus, additional assessment measures are necessary to support competent clinical decision-making.


Objectives: This study compared the diagnostic accuracy of total semantic scoring and two-dimensional bilingual coordinate scoring for Spanish-English bilingual children with and without language disorder.

Methods: Seventy-eight bilingual children with and without language disorder with 40-60% exposure to both Spanish and English participated in this study. The children completed experimental versions of the English and Spanish semantics subtests
from the Bilingual English-Spanish Assessment (BESA). Discriminant function analyses using both scoring methods were used to calculate the sensitivity, specificity, and likelihood ratios for each method.

**Results:** Both total semantic scoring and two-dimensional bilingual coordinate scoring yielded acceptable levels of sensitivity and specificity. The two-dimensional bilingual coordinate scoring method had greater accuracy overall, with a sensitivity of 93.3% and a specificity of 96.8%. Both methods yielded greater classification accuracy than if only one language had been tested.

**Conclusions:** Greater diagnostic accuracy is achieved for bilingual children when both languages are tested.

**Relevance to the current study:** Best assessment practices for bilingual children include testing in both of the child’s languages. The current study examines NWR performance in both Spanish and English to increase the classification accuracy of the task.


**Objectives:** The purpose of this study was to investigate the diagnostic accuracy of a dynamic assessment of narrative language (DYMOND) for the identification of language disorder in bilingual children.

**Methods:** The DYMOND was administered to 43 bilingual children from kindergarten to third grade and consisted of two test-teach-retest sessions. Teaching
sessions involved the teaching of story grammar and subordination for language complexity. Posttest scores, gain scores, teaching duration, and modifiability ratings were analyzed using a discriminant function analysis to determine which was the best predictor of language disorder.

Results: The overall modifiability rating was the best predictor of language disorder with excellent sensitivity (100%) and good specificity (88%) after only one session, increasing to perfect sensitivity and specificity after two sessions. Good accuracy was also found for any combination of two predictors.

Conclusions: Dynamic assessment can be an efficient and highly accurate assessment tool for the identification of language disorder in bilingual children.

Relevance to the current study: The DYMOND is used in the current study to determine participant group placements (TD or DLD).


Objectives: This article reviewed the psychometric properties of 21 preschool language assessments and investigated the classification accuracy of four assessments with the highest psychometric quality.

Methods: The manuals of 21 standardized language assessments were reviewed for ten psychometric criteria described by McCauley and Swisher (1984). Forty-one children with and without language disorder then completed four assessments that met six or more psychometric criteria.
Results: More than half of the assessments evaluated in this article met fewer than five psychometric criteria. Only five assessments met more than half of the criteria. Of the four tests that were administered, only one (the SPELT-II) resulted in acceptable classification accuracy.

Conclusions: Few standardized language assessments are of adequate psychometric quality to identify or rule out the presence of a language disorder. Even assessments that report good psychometric quality may fail to reach acceptable classification accuracy.

Relevance to the current study: Given the present weaknesses of standardized tests, additional assessment measures are needed to be used in combination with formal test results to increase confidence in diagnostic decisions.


Objectives: The purpose of this study was to develop a repetition task for young children with typically developing language abilities and to examine their performance on the task.

Methods: Sixty-six 2- to 4-year-old children completed a standardized receptive vocabulary test, followed by repetition task consisting of 18 words and 18 nonwords matched by length, prosody, and phonological structure. The children’s responses were recorded and repetition accuracy was calculated using both percent phonemes correct and whole word scoring methods.
Results: Both scoring methods yielded similar results; therefore, whole word scoring was used in the analysis for simplicity. Older children performed significantly better on the repetition task than younger children. The children overall repeated words more accurately than nonwords and decreased in accuracy with increasing nonword length. Performance on the receptive vocabulary test was significantly correlated with repetition accuracy.

Conclusions: The repetition task examined in this study can be successfully implemented with young children as a measure of their language processing ability. Whole word scoring can yield informative results and is more efficient than calculating percent phonemes correct.

Relevance to the current study: Both scoring methods discussed are investigated in the current study to determine if either result in greater diagnostic accuracy.


Objectives: This study investigated the diagnostic accuracy of norm-referenced language assessments using low cut-off points to identify language disorder.

Methods: Data from 43 norm-referenced tests were reviewed by three certified SLPs to determine mean group differences, sensitivity, specificity, and cut-off scores for each test.

Results: Analysis of the group mean differences suggested that children with language disorder often score within 1 SD of the mean and would therefore not be identified by the test as having a disorder using a low cut-off score. Nine of the 43
assessments reported sensitivity and specificity values in the test manual. Only five tests reached acceptable levels of sensitivity and specificity.

Conclusions: The results of this study contradict the assumption that children with language disorder often score significantly below their typically developing peers on standardized language assessments. Low cut-off scores are likely to under-identify children with language disorder.

Relevance to the current study: Norm-referenced assessments are often insufficient to confidently classify a child’s language ability. Additional, valid assessment measures are necessary to make accurate diagnostic decisions.


Objectives: This study investigated the relationship between language knowledge and performance on Spanish and English nonword repetition tasks in bilingual children. It also examined the relationship between phonological short-term memory, as measured by NWR, and semantic and morphosyntax abilities.

Methods: The participants of this study included 60 Spanish-English bilingual children between the ages of 4;6 and 6;5 with varying lengths of English exposure and input and output in each language. The children participated in NWR tasks using both English and Spanish nonwords, as well as semantic and morphosyntax screeners in both languages from the BESA. The children’s performance was analyzed by calculating the percentage of phonemes correct for each syllable length and the percentage of items
correct on each screener. ANOVA tests were performed to test the effects of nonword length, morphosyntax, semantics, language experience, and language exposure on NWR performance.

**Results:** Statistically significant direct effects were found for nonword length and morphosyntax on English NWR, and for nonword length, morphosyntax, and language output for Spanish NWR. Statistical significance was also observed between length of exposure to English and performance on both English and Spanish nonwords. Overall, accuracy of NWR in both languages tended to decrease as the nonwords increased in length. The children as a group performed better with Spanish nonwords than English, and children with later exposure to English generally performed better on longer nonwords than children with earlier English exposure.

**Conclusions:** Performance on NWR was determined to be significantly related to children’s cumulative language experience, with age at first English exposure exhibiting greater effects on NWR than current levels of output for each language. Longer exposure to Spanish was also suggested to improve performance on long nonwords, potentially due to the frequency of multisyllabic words in the Spanish language.

**Relevance to the current study:** While NWR is posited to be a language processing measure, this study suggests that performance on NWR tasks is influenced by language knowledge. For classification purposes, NWR tasks in both languages should be administered to account for bilingual children’s experience in each language.

Objectives: This article described two studies. The first investigated the influence of language exposure on sentence imitation (SI) and NWR performance in bilingual children. The second study examined the diagnostic accuracy of French NWR, SI, and receptive vocabulary scores in monolingual and bilingual children with and without language disorder.

Methods: In the first study, 5-year-old French-English bilingual children with varying levels of exposure to each language completed two processing-dependent tasks, NWR and SI, in French and English. Monolingual French and monolingual English speakers also completed the tasks in their own language. French-speaking participants additionally completed a receptive vocabulary assessment in French. The relationship between amount of language exposure and performance on the processing-dependent tasks in each language was analyzed using polynomial curve estimation. For the second study, monolingual and bilingual French-speaking children with and without language disorder completed NWR, SI, and a receptive vocabulary assessment in French. A one-way ANOVA analysis was used to examine differences in performance between groups. Sensitivity and specificity were then calculated at specified cut-off scores for each individual measure and combinations of measures.

Results: Significant associations were found between amount of language exposure and performance on the English NWR, English SI, and French SI tasks. Performance on the French NWR task was not significantly correlated with amount of French exposure. Of all the measures, receptive vocabulary was the most affected by language exposure. Additionally, the bilingual children’s performance on NWR with increasing word length was mostly unaffected by amount of language exposure,
suggesting that increased difficulty with longer nonwords is primarily the result of processing difficulties rather than limited exposure to the task language. Both NWR and SI yielded high levels of sensitivity for bilingual and monolingual children. Specificity for the two tasks was high for monolingual children, but much lower for bilingual children. NWR yielded greater specificity for bilingual children (79%) than SI (57%). The receptive vocabulary measure yielded poor specificity for the bilingual children.

**Conclusions:** The results of these studies indicate that processing-dependent measures are somewhat influenced by language exposure, though to a much lesser extent than knowledge-dependent measures. In particular, NWR appeared to be the least affected by language exposure and the most accurate diagnostic tool of the measures examined in this study.

**Relevance to the current study:** NWR is less influenced by language knowledge and therefore holds promise as a less-biased assessment tool for bilingual Spanish-speaking children.


**Objectives:** The purpose of this study was to further investigate the accuracy of English and Spanish NWR tasks for bilingual children and examine correlations between performance and task language.

**Methods:** One hundred eighty-seven children composing four groups (bilingual LD, bilingual TD, monolingual LD, and monolingual TD) completed a Spanish and English NWR task. Participant responses were scored by percent phonemes correct.
Results: Significant correlations were found between repetition accuracy and task language. Both mono- and bilingual LD groups performed below the TD groups on the English task. Mono- and bilingual children with LD also performed below the TD groups on the Spanish task; however, the bilingual LD group outperformed the monolingual LD group and the bilingual TD group outperformed the monolingual TD group. The English task yielded high sensitivity and low specificity for bilingual children, while the Spanish task yielded low sensitivity and good specificity.

Conclusions: The results indicate that NWR performance is influenced by exposure to the language of the task in addition to overall language ability (LD vs TD).

Relevance to the current study: The current study will investigate the same NWR tasks included in this study to further compare their diagnostic accuracy as individual and combined measures.
APPENDIX B

Institutional Review Board Approval Letter

Memorandum

To: Douglas Petersen
Department: BYU - EDUC - Communications Disorders
From: Sandee Aina, MPA, HRPP Associate Director
Wayne Larsen, MAcc, IRB Administrator
Date: March 31, 2022
IRB#: IRB2020-328
Title: Examining the Validity and Reliability of Dynamic Assessments of Reading and Language

Brigham Young University’s IRB has approved with conditions, the research study referenced in the subject heading as exempt level, categories 1 and 2. The study cannot occur until you have received specific research site approvals from the school districts in Utah and the approval of schools and districts outside of Utah.

This study does not require an annual continuing review. Each year near the anniversary of the approval date, you will receive an email reminding you of your obligations as a researcher and to check on the status of the study. You will receive this email each year until you close the study.

The study is approved as of 03/31/2022. Please reference your assigned IRB identification number in any correspondence with the IRB.

Continued approval is conditional upon your compliance with the following requirements:

1. A copy of the approved informed consent statement can be found in iRIS. No other consent statements should be used. Each research subject must be provided with a copy or a way to access the consent statement.
2. Any modifications to the approved protocol must be submitted, reviewed, and approved by the IRB before modifications are incorporated in the study.
3. All recruiting tools must be submitted and approved by the IRB prior to use.
4. Instructions to access approved documents, submit modifications, report adverse events, can be found on the IRB website, iRIS guide: https://irb.byu.edu/iris-training-resources
5. All 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. Please refer to the IRB website for more information.
## English Nonword Repetition Protocol

<table>
<thead>
<tr>
<th>Target</th>
<th>C</th>
<th>V</th>
<th>C</th>
<th>V</th>
<th>C</th>
<th>V</th>
<th>C</th>
<th>V</th>
<th>C</th>
<th>WW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) nei)b</td>
<td>n</td>
<td>ei</td>
<td>b</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) vup</td>
<td>v</td>
<td>u</td>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) ta)q3</td>
<td>t</td>
<td>ao</td>
<td>d3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) do)if</td>
<td>d</td>
<td>ci</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) te)rvak</td>
<td>t</td>
<td>ei</td>
<td>v</td>
<td>a</td>
<td>k</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6) tf)aqæg</td>
<td>tf</td>
<td>ao</td>
<td>v</td>
<td>æ</td>
<td>g</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7) vætf)æp</td>
<td>v</td>
<td>æ</td>
<td>tf</td>
<td>ai</td>
<td>p</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8) na)tf)æof</td>
<td>n</td>
<td>ci</td>
<td>tf</td>
<td>ao</td>
<td>f</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9) tf)inæta)ob</td>
<td>tf</td>
<td>i</td>
<td>n</td>
<td>ci</td>
<td>t</td>
<td>ao</td>
<td>b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10) na)tf)æve)ib</td>
<td>n</td>
<td>ai</td>
<td>tf</td>
<td>a</td>
<td>v</td>
<td>ei</td>
<td>b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11) do)ta)ovæb</td>
<td>d</td>
<td>ci</td>
<td>t</td>
<td>ao</td>
<td>v</td>
<td>æ</td>
<td>b</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12) te)rvæt)seig</td>
<td>t</td>
<td>ei</td>
<td>v</td>
<td>ci</td>
<td>tf</td>
<td>ei</td>
<td>g</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13) væt)tvæt)seip</td>
<td>v</td>
<td>er</td>
<td>t</td>
<td>a</td>
<td>tf</td>
<td>ei</td>
<td>d</td>
<td>ci</td>
<td>p</td>
<td></td>
</tr>
<tr>
<td>14) dætvæt)ævæg</td>
<td>d</td>
<td>æ</td>
<td>v</td>
<td>oo</td>
<td>n</td>
<td>ci</td>
<td>tf</td>
<td>i</td>
<td>g</td>
<td></td>
</tr>
<tr>
<td>15) nai)tf)ævævub</td>
<td>n</td>
<td>ai</td>
<td>tf</td>
<td>ci</td>
<td>t</td>
<td>ao</td>
<td>v</td>
<td>u</td>
<td>b</td>
<td></td>
</tr>
<tr>
<td>16) tævat)finæg</td>
<td>t</td>
<td>æ</td>
<td>v</td>
<td>a</td>
<td>tf</td>
<td>i</td>
<td>n</td>
<td>ai</td>
<td>g</td>
<td></td>
</tr>
</tbody>
</table>

Total Phonemes Correct: 96
Total Words Correct: 16
## APPENDIX D

### Spanish Nonword Repetition Protocol

<table>
<thead>
<tr>
<th>Target</th>
<th>c</th>
<th>v</th>
<th>c</th>
<th>v</th>
<th>c</th>
<th>v</th>
<th>c</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) du</td>
<td>d</td>
<td>u</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2) gi</td>
<td>g</td>
<td>i</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3) pe</td>
<td>p</td>
<td>e</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4) bo</td>
<td>b</td>
<td>o</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5) 'naye</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>'n</td>
<td>a</td>
</tr>
<tr>
<td>6) 'bopa</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>'b</td>
<td>o</td>
</tr>
<tr>
<td>7) 'tuno</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>'t</td>
<td>u</td>
</tr>
<tr>
<td>8) 'defu</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>'d</td>
<td>e</td>
</tr>
<tr>
<td>9) te'beta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>t</td>
<td>e</td>
</tr>
<tr>
<td>10) tso'beta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>ts</td>
<td>o</td>
</tr>
<tr>
<td>11) be'fipo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>b</td>
<td>e</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td></td>
</tr>
<tr>
<td>12) nu'tife</td>
<td>n</td>
<td>u</td>
<td>'t</td>
<td>i</td>
<td>f</td>
<td>e</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13) tʃi'nọ'teβa</td>
<td>tʃ</td>
<td>i</td>
<td>n</td>
<td>o</td>
<td>'t</td>
<td>e</td>
<td>β</td>
<td>a</td>
</tr>
<tr>
<td>14) notʃi'βeni</td>
<td>n</td>
<td>o</td>
<td>tʃ</td>
<td>i</td>
<td>'β</td>
<td>e</td>
<td>n</td>
<td>i</td>
</tr>
<tr>
<td>15) ditu'jaβu</td>
<td>d</td>
<td>i</td>
<td>t</td>
<td>u</td>
<td>'j</td>
<td>a</td>
<td>β</td>
<td>u</td>
</tr>
<tr>
<td>16) teʃo'tʃiγo</td>
<td>t</td>
<td>e</td>
<td>ʃ</td>
<td>o</td>
<td>'tʃ</td>
<td>i</td>
<td>γ</td>
<td>o</td>
</tr>
<tr>
<td>17) beteʃo'cupe</td>
<td>b</td>
<td>e</td>
<td>t</td>
<td>e</td>
<td>ʃ</td>
<td>o</td>
<td>ʃ</td>
<td>u</td>
</tr>
<tr>
<td>18) doβinu'tʃiγo</td>
<td>d</td>
<td>o</td>
<td>β</td>
<td>i</td>
<td>n</td>
<td>u</td>
<td>'tʃ</td>
<td>i</td>
</tr>
<tr>
<td>19) nitʃuto'naβe</td>
<td>n</td>
<td>i</td>
<td>tʃ</td>
<td>u</td>
<td>t</td>
<td>o</td>
<td>na</td>
<td>β</td>
</tr>
<tr>
<td>20) tebetʃi'noγe</td>
<td>t</td>
<td>e</td>
<td>β</td>
<td>e</td>
<td>tʃ</td>
<td>i</td>
<td>n</td>
<td>o</td>
</tr>
</tbody>
</table>
Nonword Scoring Guidelines
Adapted from Gray et al., 2019

First- transcribe the child’s production.
Second- scoring

1. **False starts are not errors:** omit these from the placement in the nonword sequence.

   du, du, tupwib—score as /tupwib/

   wæ, wi, wi}$/ae—score as /wi}$/ae/

2. **‘Ums’ and ‘hu’ are not errors:** omit these from the placement in the nonword sequence.

   Um, um, gɛn, um gɛnfud—score as /gɛnfud/

   Um, um, uh, wiv, um, iŋktuf—score as /wivŋktuf/

3. Line up nonword phonemes appropriately for most credit (e. g., missing syllables, phoneme addition or deletion) in the placement in the nonword sequence.

4. You may “slide” phonemes or syllables to maximize points that a child receives. Sliding a child’s response to maximize points earned is acceptable under the following conditions:

   a. Syllables must retain the order of the child’s response:

      Target: /nudɛgdɑɪŋʃup/  n u d f e g d ɑɪ n j u p
      Response: “mɛtɪdhu”  m e t I d h u

   b. If a word can be scored in more than one way, score it in a way that the child gets the most points. For example, in the example below, you would choose to score it as Option 1 because the child receives 4 points versus 3 points (Option 2).

      Target: /wɛfjoʊktuɡɪɡ/  w e f j oʊ k t u g h i ɡ
      Response: “jʌkhag”  j ʌ k h a g  j ʌ k h a g

   c. Syllables can hang off the end or beginning of the word if lining syllables up will earn the child additional points.

      Example:  Acceptable
Target: /jiktuf/  
Response: “tui”  

Example:  
Acceptable

Target: /καίμπεγ/  
Response: “dufkαιμ”  

Example:  
Acceptable

Target: /νιτʃοναβε/  
Response: “τʃινοhuhahe”  

Example:  
Acceptable  Not Acceptable

d. Syllable structure must be maintained (i.e. it is not okay to break up a CVC response structure across target syllables).

Target: /καίμτυπ/  
Response: “mup”  

5. You may “pop” out syllables or extra phonemes to maximize points. Popping a syllable/phoneme out is only okay under the following conditions:

a. If the child produces more syllables/phonemes than the number of target syllables/phonemes – pop out the excess syllable/phonemes to maximize the child’s points.

Example:  
Acceptable

Target: /jiktuf/  
Response: “jikatuf”
(note that in this case that, although the child got all of the phonemes correct, the “whole word” would not be counted as correct because of the extra syllable.

Example:  

Target:  /jitgəm/ |  j i  t  g  ə  m  
Response: “jisgrəm” |  j i  s  g(r) ə  m

6. Line up vowels with vowels and consonants with consonants.

7. Line up the child’s response from the beginning of the nonword, unless there is a phoneme similar to the one in the nonword (e.g. /b/ for /p/), in which case you may slide it over as it is the closest substitution.

Example:

Target:  / wɛfjʊktugɪg / |  w  e  f  j  ʊ  k  t  u  g  h ɪ  g  
Response: “dug” |  d  u  g