ASHA American Speech-Language-Hearing Association

JSLHR

Research Article

Identifying Language Disorder in Bilingual Children Using Automatic Speech Recognition

Nahar Albudoor[a] and **Elizabeth D. Peña**[b]

[a] Gallaudet University, Washington, DC [b] University of California, Irvine

ARTICLE INFO

Article History:
Received December 17, 2021
Revision received March 1, 2022
Accepted April 24, 2022

Editor-in-Chief: Stephen M. Camarata
Editor: Mary Alt

https://doi.org/10.1044/2022_JSLHR-21-00667

ABSTRACT

Purpose: The differential diagnosis of developmental language disorder (DLD) in bilingual children represents a unique challenge due to their distributed language exposure and knowledge. The current evidence indicates that dual-language testing yields the most accurate classification of DLD among bilinguals, but there are limited personnel and resources to support this practice. The purpose of this study was therefore to determine the feasibility of dual-language automatic speech recognition (ASR) for identifying DLD in bilingual children.
Method: Eighty-four Spanish–English bilingual second graders with (n = 25) and without (n = 59) confirmed diagnoses of DLD completed the Bilingual English-Spanish Assessment–Middle Extension Morphosyntax in both languages. Their responses on a subset of items were scored manually by human examiners and programmatically by a researcher-developed ASR application employing a commercial speech-to-text algorithm.
Results: Results demonstrated moderate overall item-by-item scoring agreement (k = .54) and similar classification accuracy values (human = 92%, ASR = 88%) between the two methods using the best-language score. Classification accuracy of the ASR method increased to 94% of cases correctly classified when test items with poorer discrimination in the ASR condition were eliminated.
Conclusion: This study provides preliminary support for the technical feasibility of ASR as a bilingual expressive language assessment tool.
Supplemental Material: https://doi.org/10.23641/asha.20249994

There is a critical need for valid language measures for the 12 million bilingual children in the United States (U.S. Census Bureau, 2019). Bilingual and other linguistically diverse children are disproportionately over- and underidentified with developmental language disorder (DLD; Morgan et al., 2015; Sullivan & Bal, 2013), an impairment in the comprehension and production of language that affects one in 10 children (Norbury et al., 2016). The consequences of DLD misdiagnosis have significant societal impact. Underidentification of DLD is associated with increased academic failure, dropout, and incarceration (Anderson et al., 2016; Thurlow & Johnson, 2011), risks that disproportionately affect persons who are minoritized (Pettit & Western, 2004; L. Wood et al., 2017). Overidentification of DLD compromises resource allocation and contributes to the increased marginalization of language minority students who receive diagnoses of DLD simply because they are bilingual.

Best practices for assessing bilingual children require evaluators to consider both languages of a child (American Speech-Language-Hearing Association [ASHA], 2021; Royal College of Speech and Language Therapists, 2021). However, only 8% of U.S. speech-language pathologists (SLPs), the practitioners qualified to diagnose DLD, are bilingual (ASHA, 2020) relative to the 22% of bilinguals in the U.S. population (U.S. Census Bureau, 2019). To address this disproportionality, it is critical to equip all SLPs with tools for assessing bilingual children, even when the SLPs are monolingual or bilingual in a language pair that differs from that of the child. Automatic speech recognition (ASR), the technology that processes human speech and converts it to text, holds promise for meeting this need. Multipurpose consumer ASR systems, such as Google Assistant and Amazon Alexa, can process up to

Correspondence to Nahar Albudoor: n.albudoor@gmail.com. ***Disclosure:*** *Elizabeth D. Peña is an author of the Bilingual English-Spanish Assessment and receives royalties. Nahar Albudoor has declared that no competing financial or nonfinancial interests existed at the time of publication.*

120 human languages and variants, and their algorithms are available for use in custom applications, whereas custom-developed ASR systems can be trained and programmed for specific use cases (Sabu & Rao, 2018). An ASR system that achieves a sufficient level of transcription agreement with proficient speakers of a target language may extend access to that language to all SLPs, enabling them to obtain information about children's skills in languages they do not speak. While this technology would not take the place of a comprehensive evaluation by a practitioner who speaks both or all languages of a child, it may serve as an indicator of that child's risk for DLD.

There is longstanding and emerging evidence that ASR is technically and practically feasible for assessing children's language skills. Dictation software, which converts speech to text, has long been used as an academic accommodation for children with special education needs (Bolt & Thurlow, 2004; Thompson et al., 2002) and can achieve a mean word-level accuracy of 87% (MacArthur & Cavalier, 2004). Newer educational software using ASR and artificial intelligence conducts more advanced conversational exchanges with children as successfully as human interlocuters, including delivering prompts such as story comprehension questions, processing children's oral responses to the prompts, and responding appropriately (Xu et al., 2021). Still, ASR has yet to be employed in the assessment of children's language skills and the extent to which ASR can be used to identify DLD is currently unknown. Thus, the purpose of this study was to provide an early indication of the extent to which ASR may be used for DLD identification. A researcher-developed ASR application was employed to transcribe the responses of Spanish–English bilingual second graders who completed a validated morphosyntax assessment task in Spanish and English. Children were in two groups: those with DLD and those with typical language development (TD). Children's ASR-transcribed responses were programmatically scored and analyzed for agreement with human scorers and for DLD classification accuracy to determine the feasibility of ASR for identifying DLD in bilingual children.

Identifying DLD in Bilingual Children

Despite a consensus on the importance of dual-language testing of bilingual children (Anaya et al., 2016; ASHA, 2021; Peña & Bedore, 2011; Peña et al., 2016, 2020; Royal College of Speech and Language Therapists, 2021), there are significant barriers to its implementation that can contribute to the over- or underidentification of DLD in this population. To collect information about two languages, evaluating SLPs either must be proficient in both languages or must establish a method of obtaining information about the language they do not speak. Among the small percentage of SLPs who are bilingual, challenges that include lack of time and access to assessment materials/training are cited as common barriers to bilingual testing (Arias & Friberg, 2017). Among SLPs who are not bilingual, methods such as collaborating with interpreters also have challenges. For example, Saenz and Langdon (2019) found that, of the 208 California SLPs they surveyed who reported previously working with interpreters, most reported that they had experienced instances when they needed to work with an interpreter but could not. This was most commonly due to an inability to find an interpreter (69%), uncertainty about the interpreter's training (26%), lack of necessary assistance from the interpreter (23%), or lack of monetary support from an employer (16%). Such challenges represent barriers to the timely and accurate diagnosis of DLD among bilinguals that increases their risk for DLD misdiagnosis.

Automatic Assessment to Expand Access to Accurate Identification

Recent work has explored the use of automatic language tasks as a potential alternative to person-administered dual-language assessment. For example, de Villiers et al. (2021) reported on the development of the Quick Interactive Language Screener: English–Spanish (QUILS: ES), an electronic language screening instrument that automatically administers and scores receptive vocabulary, syntax, and processing tasks in both English and Spanish using a touchscreen tablet interface. The QUILS: ES results of Spanish–English bilingual children aged 3 to 5;11 (years; months) were significantly correlated with their results on the English ($r = .693$, $p < .001$) and Spanish ($r = .449$, $p < .002$) Preschool Language Scales–Fifth Edition, providing evidence for the instrument's concurrent validity. Furthermore, the instrument produced good internal consistency (English = .89, Spanish = .85) and test–retest reliability (.89), indicating good fit of the test items with the measurement constructs and moderately high correlations between measurement occasions, respectively. Similarly, pilot analyses from the study of Pratt et al. (2022) explored the feasibility of a remote, computer-administered English and Spanish language assessment protocol that employed prerecorded test items to measure children's oral language skills. In response to receptive test items, children listened to the prompts and clicked on their responses from picture arrays, which were automatically scored. In response to expressive test items, children listened to the prompts, answered verbally, and their responses were manually scored by the evaluators. Results from the split-half administrations demonstrated a significant correlation ($r = .979$, $p < .01$) between children's results in the remote condition and their results in the in-person condition, demonstrating the instrument's reliability across administration conditions.

These previous studies have represented an emerging area of research in bilingual DLD assessment that holds considerable promise for extending access to multiple languages to all SLPs. A primary limitation of this work, however, is that automatic scoring is limited to receptive test items. Expressive tasks, however, represent a core aspect of bilingual DLD assessment. Among bilingual DLD identification measures, all measures with fair to good classification accuracy (i.e., sensitivity and specificity above 80%) reviewed by Brinson et al. (2020) contained expressive production tasks or test items. As such, the current evidence provides preliminary technical and practical support for the automatic scoring of receptive language tasks, but further research is necessary for establishing the automatic scoring of expressive language tasks. Such work is a critical step toward the development of instruments that are fully automatically administered and scored and is therefore the focus of this study.

Automatic Recognition of Children's Speech

There is considerable evidence that ASR is technically viable for processing children's speech for the purposes of expressive language assessment. Dictation through ASR software has existed as an academic accommodation for children receiving special education services from as early as 1997 when Dragon Systems released the first computer software that converted connected speech from audio to text (Dragon Systems, Inc, 1997; Thompson et al., 2002). An early study of 14-year-olds with and without learning disorders (LDs) who completed a sentence probe task using the Dragon: Naturally Speaking software demonstrated that the software produced an overall word accuracy rate of 87%, with no significant difference in accuracy between the LD and non-LD groups (MacArthur & Cavalier, 2004). These results provided initial evidence that ASR systems could capture the speech of children, even those with special needs. More recently, researcher-developed ASR systems have achieved even higher accuracy rates with younger children including those with speech production deficits. Hair et al. (2019) tested multiple trained ASR models for their accuracy for analyzing the speech of children diagnosed with speech sound disorders at the single-word level. The best-performing ASR model achieved a mean 90% accuracy. At the sentence level, Sabu and Rao (2018) developed an ASR system that achieved a word error rate (i.e., the percentage of recognized words containing substitutions, deletions, or insertions) of just 3.44% when it was used to analyze the speech of 20 students between the ages of 10 and 14 years who completed a sentence oral reading task. These findings indicate that both commercial and custom ASR applications can achieve high accuracy and low error rates when processing children's speech even when it contains production errors. Although what constitutes "high-enough" ASR transcription accuracy is ultimately usage dependent, adults with normal hearing achieve an average word recognition accuracy of approximately 95% when listening to other adults and the signal-to-noise ratio is above −5 dB (Spille et al., 2018). As such, some of the child ASR models in the reviewed studies achieved word accuracy rates near or on par with adult human listeners of other adult speakers.

ASR for Bilingual Language Assessment

The research on the technical feasibility of ASR for recognizing children's speech indicates that ASR-embedded applications can be used to process and score test items for the purposes of language evaluation. However, this research does not demonstrate the extent to which this applies to languages other than English and for children with DLD. A pilot analysis of Spanish–English bilingual second graders' test audio recordings demonstrated initial evidence toward this aim (Albudoor et al., 2019). Employing the Google Cloud speech-to-text application programming interface (API), we analyzed the audio recordings of 20 Spanish–English bilingual second graders, 10 of which with confirmed diagnoses of DLD and 10 of which with TD who were matched by age, sex, maternal education, and percent current English language exposure to peers in the DLD group. The children's audio recordings contained their responses to test items from the morphosyntax and semantics subtests of the Bilingual English-Spanish Assessment–Middle Extension (BESA-ME; Peña et al., 2012a) and the narrative comprehension scale of the Test of Narrative Language (TNL; Gillam & Pearson, 2004; Gillam et al., 2010) in both English and Spanish. The ASR's overall item-level agreement with human scoring of these measures was 81% for English items and 84% for Spanish items, indicating that the technology demonstrated moderate agreement with human scorers for assessing both Spanish and English test items. The findings provided evidence for the use of ASR to assess the language skills of Spanish–English bilingual children in that they demonstrated that a reasonable degree of scoring agreement could be achieved with ASR. However, there were two limitations to these pilot analyses. First, scoring agreement was calculated using simple percentage agreement by item, which does not account for the possibility of chance agreements like more sophisticated metrics of interrater reliability, such as Cohen's kappa (Cohen, 1960). Second, the classification accuracy of the measure (i.e., the degree to which it correctly classified true DLD and TD cases) was not determined as it was not the aim of the study. Establishing classification accuracy is a necessary step toward determining a measure's feasibility

for diagnostic purposes. While high human–ASR scoring agreement may suggest classification accuracy like the original measure's, ASR classification may reveal advantages and disadvantages specific to the technology and must therefore be independently confirmed. Thus, this study aimed to extend previous work to further explore human–ASR scoring agreement and to determine the extent to which an ASR measure can independently and accurately classify children with and without DLD.

Research Aims

This study aimed to determine the scoring agreement and classification accuracy of a Spanish–English expressive morphosyntax task—transcribed using ASR technology and scored programmatically—to provide evidence for its technical feasibility as an assessment tool. Specifically, the first research aim was to determine the item-level agreement between children's original (i.e., human-scored) scores on a Spanish–English bilingual morphosyntax measure and their scores when ASR transcripts were used to score the same assessment. The second research aim, pertaining to DLD identification, was to determine the degree to which the same morphosyntax measure, analyzed using the ASR transcription and scoring procedure, accurately classified children with their original TD or DLD classifications. All analyses were conducted on existing assessment data drawn from the study of Peña et al. (2010).

Method

Participants

Participants were 84 Spanish–English bilingual second graders with (n = 25) and without (n = 59) DLD whose data were drawn from 334 participants in the longitudinal phase of a study of bilingual DLD (Peña et al., 2010). Eighty-four children were included in the current analysis because they (a) completed the morphosyntax subtest of the BESA-ME Field Test (Peña et al., 2012a) in both English and Spanish at second grade and (b) had complete audio recordings. Second grade was selected as the analysis year as this was the grade with the greatest number of participants, yielding the largest DLD group for classification analyses. Table 1 shows the participant demographics. The TD and DLD groups did not significantly differ in age, sex, maternal education, first English exposure, or current English exposure. The primary Spanish dialect spoken by children in this study was Mexican (88%), followed by "other" (2%) and Salvadorean (1%); a Spanish dialect was not reported for the remaining 8% of children. The primary English dialects spoken by children in this study were Standard American English (40.5%) and Southern English (36%), followed by "other" (2%); an English dialect was not reported for the remaining 21.5% of children.

DLD Classification

Identification of DLD was conducted as part of the larger Peña et al. (2010) study and used a protocol that required converging evidence across multiple indicators. Specifically, children were classified with DLD if they met four of the five following indicators of impairment:

1. Parent or teacher concern rating (as measured by the Inventory to Assess Language Knowledge; Peña et al., 2018) below 4.2 (out of 5) in both English and Spanish;
2. BESA-ME Field Test morphosyntax score lower than 1 *SD* below the normative mean in both English and Spanish;
3. BESA-ME Field Test semantics score lower than 1 *SD* below the normative mean in both English and Spanish;
4. BESOS composite score lower than 1 *SD* below the normative mean in both English and Spanish; and/or
5. Test of Narrative Language (Gillam & Pearson, 2004; Gillam et al., 2010) composite score lower

Table 1. Participant demographics.

Variable	TD	DLD	Total	t/χ^2	p
N	59	25	84		
Age (in years) – *M (SD)*	7.9 (0.3)	7.9 (0.4)	7.9 (0.4)	0.222	.53
Sex (% female)	47%	40%	45%	0.394	.53
Maternal education[a]	2.5 (1.5)	2.3 (1.5)	2.5 (1.5)	–0.790	.43
Age of first English exposure (in years) – *M (SD)*	2.7 (1.6)	3.4 (1.44)	2.9 (1.6)	2.034	.05
Percent current English Input/output[b] – *M (SD)*	43.0 (14.4)	38.2 (13.3)	41.5 (14.9)	–1.440	.16

Note. TD = typical language development; DLD = developmental language disorder.

[a]Hollingshead (1975) score, where 1 = less than 7th grade, 2 = junior high (9th grade), 3 = partial high school (10th or 11th), 4 = high school graduate, 5 = partial college (at least one year), 6 = college education, and 7 = graduate degree. [b]Derived by averaging the percentage of weekly hours spent hearing English with the percentage of weekly hours spent speaking English (relative to Spanish).

than 1 *SD* below the normative mean in both English and Spanish.

Children were classified as TD if they demonstrated three or fewer of these indicators. Children were excluded from the larger study and, therefore, the present analyses if they presented with a history of focal brain injury, autism spectrum disorder, intellectual impairment, socioemotional disorder, or hearing loss.

Materials

Reference Measures

The following reference measures were used to identify indicators of impairment for children's original DLD diagnoses. Based on their results on these measures, children were classified with DLD if they met four of the five classification indicators listed above.

Bilingual English Spanish Oral Screener

The Bilingual English Spanish Oral Screener (BESOS; Peña et al., 2012b) is a language screening for Spanish–English bilingual children between prekindergarten and third grade. Preliminary norming for the BESOS demonstrates sensitivity of .80 to .93 and specificity of .92 to .94 (depending on the age group) for identifying language disorder (Peña et al., 2018), which is above the .80 cutoff Plante and Vance (1994) designated as "fair" for identifying language disorders in children. The BESOS contains one semantics subtest and one morphosyntax subtest in English and Spanish. The semantics subtests measure children's depth and breadth of word knowledge through structures such as functions, definitions, and analogies. The morphosyntax subtests measure children's morphological and syntactic structures through cloze and sentence repetition items. In English, structures include possessive *'s,* regular/irregular past tense, and passives. In Spanish, structures include object clitics, relative clauses, and subjunctives.

Inventory to Assess Language Knowledge

The Home and School Inventory to Assess Language Knowledge (ITALK; Peña et al., 2018) measure parent- and teacher-reported child language knowledge, respectively. Parents/caregivers and teachers rate children's vocabulary, speech, sentence production, grammar, and comprehension skills on a scale from 0 to 5 for both Spanish and English. Respondents receive descriptors and examples for each point on the scale in order to select the score that best represents the child's skills. The five scores are then averaged to yield one Home ITALK and one School ITALK score for each language that falls between 0 and 5, with 0 representing no skills and 5 representing extensive skills.

BESA-ME Field Test

The BESA-ME Field Test (Peña et al., 2012a) is a dual-language measure intended for use with Spanish–English bilingual children between the ages of 7;0 and 11;6 (see Bedore et al., 2018). The Spanish and English semantics subtests measure semantics breadth and depth through receptive and expressive item types that evaluate a child's ability to identify word functions, categories, definitions, characteristic properties, analogies, similarities and differences, and associations. The English morphosyntax subtest examines possessive *-s*, third-person singular, regular past tense, plural nouns, present/past auxiliary + progressive *-ing*, copula negatives, and passives. The Spanish morphosyntax subtest examines articles, present progressive verbs, direct object clitics, and subjunctives. The morphosyntax subtests are divided into cloze and sentence repetition sections. Test items from the cloze task require children to expressively complete sentences with words or phrases containing target morphosyntactic forms. Test items from the sentence repetition task require children to verbally repeat full sentences containing target morphosyntactic forms. Children are assessed both on their ability to repeat the whole sentence (verbatim scoring) and on their ability to repeat individual word and phrase targets from the sentence (target scoring). Preliminary classification analyses for the BESA-ME Field Test demonstrate sensitivity of 1.0 and specificity of .87 to .95 (depending on the age group) using the best-language morphosyntax and semantics composite.

TNL

The TNL English (Gillam & Pearson, 2004) is a published, norm-referenced measure of children's narrative language skills for children between the ages of 5;0 and 11;11. The TNL Spanish (Gillam et al., 2010) is an experimental test identical in structure to the TNL English and for which preliminary norming has been conducted. The tests consist of two scales: Narrative Comprehension and Oral Narration. The Narrative Comprehension scale requires children to answer comprehension questions about three oral stories. The Oral Narration scale requires children to retell one oral story using no visual prompts and tell two oral stories, one while viewing a sequence of five pictures and another while viewing a single picture. For the TNL English (Gillam & Pearson, 2004), Hispanic children make up 12% of the normative sample and the measure has been validated for use with bilingual children (Gillam et al., 2013). Sensitivity and specificity for confirming the presence or absence of language disorder are .92 and .87, respectively. For the TNL Spanish (Gillam et al., 2010), preliminary data based on 216 children suggest

alpha levels (i.e., internal consistency) of .89 and .93 for the Narrative Comprehension and Oral Narration scales, respectively (Peña et al., 2020). Furthermore, data from a subset of 90 children showed that children with TD receive significantly higher raw scores on the TNL Spanish subtests (M = 8.6) than children with DLD (M = 4.4).

Index Measure

A composite measure consisting of a subset of BESA-ME morphosyntax items served as the index measure for evaluating the feasibility of ASR for DLD identification in this study. This measure was selected because (a) it elicits expressive productions and is therefore a candidate for ASR scoring and (b) in an analysis of second and fourth grade Spanish–English bilinguals, Peña et al. (2020) demonstrated that the BESA-ME morphosyntax accounted for the most variance in discriminating between TD and DLD second graders (the age group of interest in this study), over and above the BESA-ME semantics and TNL. To ensure test items' utility for disorder identification purposes and with children with varying degrees of English language exposure, items in the composite measure that met the following criteria were included in the index measure:

1. an item discrimination index at or above .30 between TD and DLD children (Ebel & Frisbie, 1986), calculated per D. A. Wood (1960);
2. the item mean of the TD group was at least .3 point higher than the item mean of the DLD group; and
3. the item mean was at least .30 for at least two of three language exposure profiles (English-dominant [60% or more current English exposure], Spanish-dominant [40% or less current English exposure], or balanced [40%–60% current English exposure]).

The item-level data of all children who completed the BESA-ME morphosyntax subtests during the larger longitudinal study were analyzed to identify the test items that met these criteria. This sample included 283 children (TD = 237, DLD = 46) who completed the English morphosyntax subtest during at least one test year, contributing an average of 2.3 datapoints per English item (i.e., longitudinally on the same test item), and 252 children (TD = 212, DLD = 40) who completed the Spanish morphosyntax subtest during at least one test year, also contributing an average of 2.3 datapoints per Spanish item. This resulted in a total item set that consisted of 646 responses per English item and 569 responses per Spanish item. Children's classifications were determined using the DLD identification protocol of the original study, as discussed above. At each time point, children's documented current language exposure was used to determine their language exposure profiles. This procedure resulted in 34 English (cloze = 18, sentence repetition = 16) and 27 Spanish (cloze = 5, sentence repetition = 22) items from the original 102 English and 108 Spanish BESA-ME morphosyntax items. Discrimination indices ranged from .43 to .78. The average discrimination index for the English items was .67 (SD = 0.09), whereas the average discrimination index for the Spanish items was .61 (SD = 0.09).

ASR Application

The ASR application used to transcribe children's test responses for the current analyses was a researcher-coded Python program that employed version 1 of the Google Cloud nonstreaming REST speech-to-text API (Google, 2020). REST API is a programmable algorithm developed by Google that asynchronously converts human speech to text across 125 languages and language variants. While it is commercially available in multipurpose consumer devices and software, it is also available for use in custom applications at a cost-per-minute basis. To include the REST API in a custom program, a JavaScript object notation (JSON) access token associated with a Google Cloud account is written into a developer's custom code using the programming language of choice (Python was used in this study). The access token then allows the custom program to send audio data to the Google server, where it is converted to text and returned to the user. As there are multiple language and model options, the code is programmable for the target language and target type of model. In this study, the *en-US* (United States English) and *es-US* (United States Spanish) "command and prompt" models, which Google specifies are suitable for analyzing short segments of speech (Google, 2020), were employed. Given that these models are language specific, all input is treated as target-language input regardless of the actual language used (e.g., Spanish input is treated as English input when the *en-US* model is in use).

Procedures

Data Collection

The present analyses were conducted on children's existing language assessment audio recordings, collected during the Peña et al. (2010) longitudinal study. Approval to recruit and consent participants was obtained from the institutional review board of The University of Texas at Austin (study 2009-11-0110). During the screening phase of the study, children completed the BESOS in English and Spanish. Trained Spanish–English bilingual research assistants administered the BESOS to children individually at their schools. Generally, all subtests and languages of

the BESOS were completed within a single 30-min session. Children's responses were recorded on paper test forms that were later digitally scanned and uploaded to a secure file server.

During the testing phase of the study, which began 1 year after the screening phase, children completed a battery of language and cognition measures once per year for up to 4 years (see Peña et al., 2020, for detailed descriptions of these measures). Bilingual research assistants administered all test measures to children individually at their schools in a quiet space. Testing was completed over three to six sessions that were 30 min to an hour in length. Children's responses were manually recorded on paper test forms and audiorecorded using Zoom H2n Handheld SD Recorders in .mp3 320 kbps acg2 (for speech) mode. The scanned paper test forms and digital audio recordings were later uploaded to a secure file server. Parents and teachers completed the ITALK in person or over the phone.

Data Analysis

Audio Recording Processing

There were two existing audio recordings per child, one from each of their BESA-ME morphosyntax testing sessions (English and Spanish). Children's responses to the target test items were extracted from the longer audio recordings using Audacity Version 2.3.2 (Audacity Team, 2020). This yielded an individual audio recording for each test item response to simulate the length of the responses consistent with a conversational agent employing ASR during testing. The segmented audio files were saved in the .wav file format at the original 16000-Hz sampling rate and monosignal.

ASR Transcription

To convert children's audio files to ASR-transcribed item responses, the researcher-coded ASR speech-to-text Python program: (a) extracted the audio recording file from a local directory; (b) scanned the file name for the target language (English or Spanish); (c) converted the entire audio response to text using the target language speech-to-text transcription model, generating up to four transcription alternatives; and (d) outputted the transcription alternatives to a .csv file, with one row representing one audio file.

Scoring

Children's original scores on the test items, scored by human evaluators during testing, were drawn from the existing data set. Scoring reliability was performed on 10% of the samples from the original longitudinal study and yielded an average 99.8% interrater reliability, ensuring the reliability of the human evaluators. To determine children's ASR-transcribed scores for each item, an R script programmatically compared children's transcripts to a set of target responses derived from the BESA-ME morphosyntax record protocols. That is, each item had a narrow set of acceptable target responses that allowed for variation in the production of the target form (e.g., "where is…" and "can you…" when the target was question inversion) or shorthand usages (e.g., "cause" for *because*). Children were assigned a score of 1 on a test item if the ASR-transcribed response included an acceptable target response for that item across any of the four transcription alternatives. Otherwise, they received a score of 0 on that item. See Supplemental Material S1 for a sample item, its programmed acceptable responses, and a set of ASR transcription alternatives for a participant's response to this item.

Results

Scoring Agreement

The first aim of this study was to determine the item-level agreement between children's human-scored test scores and their ASR test scores on the subset of BESA-ME morphosyntax items. There were 5,124 total item-level responses. To identify scoring agreement, the Cohen's kappa coefficient was calculated overall, by item, by test language, and by disorder classification. Per Cohen (1960), kappa coefficient values of ≤ 0 = no agreement, .01–.20 = none to slight, .21–.40 = fair, .41–.60 = moderate, .61–.80 = substantial, and .81–1.00 = almost perfect agreement. The overall item-level agreement across test languages and classifications was .54, indicating moderate overall scoring agreement between the human and ASR scores. Item-by-item, agreement ranged from slight (with a minimum k = .08) to almost perfect (with a maximum k = .85), indicating substantial variation in agreement across test items.

There were also variations in agreement between test languages and classifications. To determine whether these variations were substantial, we evaluated the overlap between the 95% confidence intervals (CIs) of the kappa coefficients. There was no overlap in the English and Spanish confidence intervals, with results indicating that the Spanish items yielded higher agreement (k = .62, 95% CI = [.59–.65]) than the English items (k = .47, 95% CI = [.44–.50]). However, agreement in both languages was at least moderate. There was also no overlap in the TD and DLD confidence intervals, with results indicating that responses by children with TD (k = .52, 95% CI = [.50–.55]) yielded higher agreement than responses by children with DLD (k = .36, 95% CI = [.31–.41]), who yielded fair agreement. The difference between TD and

DLD agreement indicated a potential risk to the classification accuracy of the ASR measure.

To further explore the relationships between the human- and ASR-scored items, the item discrimination indices of the ASR-scored items were calculated. These were significantly and positively correlated with the human-scored item discrimination indices, $r = .45$, $p = .01$, indicating their concurrent validity, but were lower overall, ranging from .03 to .63. The average discrimination index for the English items was .30 ($SD = 0.12$), whereas the average discrimination index for the Spanish items was .35 ($SD = 0.13$). Paired t tests comparing the human and ASR discrimination indices confirmed that the ASR indices were significantly lower than the human-scored indices, $t = -9.689$, $p < .001$, with a mean difference of −.14. Furthermore, discrimination indices were significantly and positively correlated with items' Cohen's kappa coefficients, $r = .45$, $p = .0003$, indicating that human–ASR item agreement was positively associated with the ASR discrimination index of that test item. These findings suggested that the ASR classification analyses would yield similar accuracies to the human-scored classification analyses but that some items with lower scoring accuracies and/or discrimination indices in the ASR-scored condition may negatively impact the ASR results.

Classification Accuracy

The second aim of this study was to determine the classification accuracy of children's ASR test scores, that is, the extent to which ASR scores accurately grouped children into the DLD and TD groups. Children's existing DLD classifications were used as the reference for examining classification accuracy. Classification analyses were conducted in three stages. First, because this study analyzed a subset of items from the BESA-ME Morphosyntax, the classification accuracy of the human-scored item subset was established (i.e., the human condition). Second, the classification accuracy of the ASR-transcribed and programmatically scored item subset was determined using the same item set as the human-scored condition (i.e., the ASR condition). Third, to explore whether the classification accuracy of the ASR condition could be improved by culling test items with poorer agreement, follow-up classification analyses were conducted (i.e., the follow-up ASR condition). In keeping with prior research, children's best-language scores (the higher percentage correct score of the two languages) were entered into all classification analyses. Classification analyses were conducted using receiver operating characteristic (ROC) curve analyses. The ROC curve analyses identified the thresholds (i.e., the percentage correct scores that maximized sensitivity and specificity, serving as the optimal cut point for discriminating between children with and without DLD) and the classification metrics associated with each threshold. The results of the ROC curve analyses are shown in Table 2.

Human Condition

At an optimal cut point of 54%, the human condition (Table 2, left column) yielded adequate sensitivity (88%) and good specificity (93%) for identifying DLD, per Plante and Vance (1994), with an overall 92% of cases correctly classified. This indicated that the set of test items could be used to classify children with and without DLD and therefore served as a robust baseline from which to analyze the ASR condition. Among children with DLD, 56% had a Spanish best-language score, 40% had an English best-language score, and 4% achieved the same score in Spanish and English. Among TD children, 63% had a Spanish best-language score, 36% had an English best-language score, and 1% achieved the same score in both languages.

ASR Condition

With the same set of test items as the human condition, at an optimal cut point of 39%, the ASR condition (Table 2, middle column) yielded the same level of sensitivity as the human-scored condition (88%). However, specificity dropped to adequate (88%), with an overall 88% of cases correctly classified. This indicated that, using the same item set, the ASR condition identified positive cases similarly to the human condition but yielded more false positives among the negative cases. Among children with DLD, 80% had a Spanish best-language score, 16% had an English best-language score, and 4% achieved the same score in Spanish and English. Among TD children, 73% had a Spanish best-language score and 27% had an English best-language score, with none achieving the same score in both languages. This finding was notable because the proportion of children with Spanish best-language

Table 2. Receiver operating characteristic curves on children's test scores across scoring methods (DLD = 25, TD = 59).

Variable	Human	ASR	Follow-Up ASR
Optimal cut point (%)	54	39	38
Accuracy (%)	92	88	94
Sensitivity (%)	88	88	84
Specificity (%)	93	88	98
True positives (*n*)	22	22	21
False negatives (*n*)	3	3	4
True negatives (*n*)	55	52	58
False positives (*n*)	4	7	1
Positive likelihood ratio	12.98	7.42	49.56
Negative likelihood ratio	0.13	0.14	0.16

Note. DLD = developmental language disorder [positive cases]; TD = typical language development [negative cases]; ASR = automatic speech recognition.

scores was higher in the ASR condition than the human condition for both DLD and TD children. This suggested that, when scored using ASR, the Spanish test items yielded a higher average than the English test items.

Follow-Up ASR Condition

Given that there were variations in item scoring agreement, a follow-up classification analysis was conducted to determine whether the ASR classification accuracy could be improved. An item selection procedure was conducted, uninformative test items were culled, and classification analyses were repeated on a new shorter item set. Only ASR items that were likely to increase classification accuracy were retained in this shorter item set. Specifically, we retained only ASR-scored items with a discrimination index at or above .3, a DLD–TD mean difference at or above .3, and a mean score at or above .3 for two of the three language exposure groups (the same item selection criteria used to construct the index measure). This yielded 15 English items (cloze = 9, sentence repetition = 6) and 17 Spanish items (cloze = 3, sentence repetition = 14). Notably, this procedure culled a combination of cloze and sentence repetition items and the proportion of item types in the follow-up ASR condition (English: cloze = 60%, sentence repetition = 40%; Spanish: cloze = 18%, sentence repetition = 82%) remained similar to the proportion of item types in the index measure (English: cloze = 53%, sentence repetition = 47%; Spanish: cloze = 18.5%, sentence repetition = 81.5%). This finding suggested that the ASR condition did not favor a particular item type. However, all sentence repetition items requiring a verbatim response were culled, indicating that this method of sentence repetition scoring did not produce adequate differences between TD and DLD children in the ASR condition.

At an optimal cut point of 38%, the follow-up ASR condition (Table 2, right column) yielded lower yet adequate sensitivity (84%) than both the human and ASR conditions, but specificity was almost perfect (98%). The proportion of DLD and TD children achieving a Spanish best-language score was identical to the ASR condition, demonstrating that item selection did not improve the ASR bias toward Spanish test items. However, the follow-up ASR condition was associated with the highest positive likelihood ratio (49.56) among the three conditions, indicating that it was associated with the highest likelihood that a positive result (i.e., a DLD case) was true (Dollaghan, 2007). The negative likelihood ratios were similar among the three conditions (.13–.16), indicating that they yielded similar likelihoods that a negative result (i.e., a TD case) was true. Additionally, of note was that the optimal cut points of both ASR conditions were 15%–16% lower than the cut point of the human condition. This indicated that a lower percentage correct score discriminated between children with and without DLD when ASR was used to transcribe and score their test responses.

Discussion

This study presents preliminary evidence for the technical feasibility of ASR as a bilingual expressive language assessment tool. The dual-language morphosyntax assessment responses of Spanish–English bilingual second graders with and without confirmed diagnoses of DLD were used to develop a bilingual English–Spanish index measure with high classification accuracy when scored by a human examiner. Children's audiorecorded responses to the items on this measure were transcribed by a researcher-developed ASR application and programmatically scored. The ASR measure achieved moderate item-by-item scoring agreement with the human-scored measure overall and agreement was significantly associated with item discrimination indices, indicating that higher agreement would yield improved classification. Despite variability in scoring agreement between test items, test languages, and disorder classifications, the classification accuracy values of the human and ASR conditions differed by just four percentage points (92% and 88% of cases classified correctly, respectively), with the ASR measure yielding the same sensitivity but lower specificity. When the ASR-scored test items were further narrowed by retaining only those with adequate or higher discrimination in the ASR condition, accuracy rose to 94% of cases classified correctly using the best-language percentage correct. This increase was related to improved specificity at the cost of sensitivity. Overall, these findings demonstrated that an identical ASR adaptation of an existing expressive morphosyntax measure achieved the same identification of DLD but slightly lower identification of TD but that the ASR test item set could be manipulated to increase specific classification metrics. Although this study did not explore the practical feasibility of automatic administration, the present findings provide evidence for the technical feasibility of ASR transcription and scoring for DLD identification.

The Feasibility of Automatic Expressive Language Assessment

The current findings suggest that children's expressive language skills (i.e., their verbal responses) in more than one language can be automatically evaluated. Specifically, we extend the works of de Villiers et al. (2021) and Pratt et al. (2022), who demonstrated the validity of automatic receptive language tasks (i.e., listening to prompts and clicking/tapping the correct responses) in two

languages. In this study, children's verbal responses to English and Spanish expressive test items were successfully automatically transcribed and scored, yielding adequate to good classification accuracy. Together with the previous research, these findings provide evidence for the technical feasibility of assessment instruments that automatically score language tasks across both the expressive and receptive modalities for DLD identification purposes.

There are two key contributions to the field of child language assessment associated with this outcome. First and more broadly, this study shows that ASR scoring of children's morphosyntax skills is viable within languages, suggesting that a range of measures can employ this technology. That is, scoring agreement was at least moderate for both the English and Spanish items. These results suggest that single-language English or Spanish measures and/or measures can be scored using ASR. This is an important contribution in that it supports a path toward the broad adoption of automatic DLD assessment instruments, which has the potential to improve the efficiency and accuracy of language assessment practices for all children, not only those who are bilingual.

A second key contribution and one more specific to the current aims is that this study supports a novel method for SLPs to assess languages they do not speak validly and reliably. Automatic dual-language assessment may reduce the reliance on bilingual SLPs and interpreters, who can be inaccessible or who may not have the resources to conduct such assessments (e.g., Arias & Friberg, 2017; Saenz & Langdon, 2019). This would allow all SLPs to collect information about both languages of a bilingual child. Used in conjunction with other critical pieces of assessment information, including language sampling, dynamic assessment, and parent/teacher interviews (see Anaya et al., 2016, and Castilla-Earls et al., 2020, for more information about converging evidence frameworks), this information can support more accurate DLD diagnosis among bilinguals, reducing the risk of the endemic inequities in educational access often faced by this population. Alternatively, automatic dual-language assessment tools may be employed as screening measures, serving as early indicators of risk for DLD.

An important consideration is that both aforementioned contributions are conditional upon SLPs' adoption of such tools, but the current evidence suggests that automatic assessments are likely to be adopted. In addition to the technical feasibility demonstrated in this study, there is emerging evidence that expressive language tasks can be automatically administered, with children as young as three successfully engaging in these tasks for as long as 30 min (Xu et al., 2021; Yeung et al., 2019). Additionally, while this study did not explore SLPs' attitudes about the adoption of ASR for language assessment, a significant factor predicting SLPs' use of clinical technologies is whether the technologies enable them to accomplish tasks more quickly and effectively (Albudoor & Peña, 2021; Boster & McCarthy, 2018). Finally, similar tools are prevalent in K-12 education, suggesting that their adoption and implementation is likely. All three of the most common K-12 English language proficiency measures in the United States—ACCESS, ELPA21, and ELPAC—are electronically administered and partially automatically scored on desktop or laptop computers (Kim et al., 2020). These tools evaluate children's listening, speaking, reading, and writing skills using tasks very similar to those employed by DLD identification instruments and have been adopted by 49 of the 50 U.S. states. While the mentioned proficiency measures do not yet automatically score children's verbal or written expressive responses, test development companies are now trialing automated speech scoring systems for child language proficiency measures such as the Test of English as a Foreign Language Junior (Evanini et al., 2020). Together, these findings indicate that more widespread implementation of automatic language assessments employing ASR is likely to occur and that SLPs are likely to adopt such technologies if they are available and effective, further confirming the importance of providing empirical support for their use in DLD identification.

Implications for Future Automatic Language Assessment

This study provides implications for the automatic assessment of English and Spanish skills for the purposes of DLD identification, providing four considerations for automatic assessment. First, this study established that there was some cost to TD–DLD discrimination associated with ASR scoring but that the original (conservative) item selection procedure prevented the ASR classification accuracy from dropping substantially. The individual discrimination indices of the test items fell significantly in the ASR condition compared to the human-scored condition, but the overall classification accuracy of ASR was only four percentage points lower and was able to be increased in follow-up analyses. Even when test items were culled in the follow-up analyses, there remained enough items to repeat the classification analyses and yield good classification accuracy. These findings suggested that it was important to begin with a larger but highly robust item subset as the index measure, as some items in the ASR condition were candidates for elimination due to inadequate discrimination indices. These findings also highlighted the potential benefits and costs of using existing validated test items. In this study, item selection for the ASR measure relied on deriving item discrimination indices from a large existing sample of children. This facilitated the efficient formulation of the ASR item subset. However, developing and evaluating test items directly for use with ASR

technology may allow test developers to bypass the drop in classification accuracy associated with paper-to-ASR adaptations. Furthermore, test development that is specific to ASR may support the validation of measures in languages for which validated assessments are currently not available.

Second, the current findings highlighted how ASR classification accuracy could vary in different directions from human-scored classification accuracy. In the first ASR classification analysis, within which all items were tested, classification accuracy dropped in specificity (i.e., the measure's ability to identify true negative [TD] cases) but not in sensitivity (i.e., the measure's ability to identify true positive [DLD] cases). In other words, the ASR condition falsely flagged more children as DLD. This suggests that, all things being equal, children are more likely to fail an ASR-scored measure compared to a human-scored measure. However, the reverse pattern was observed when the ASR measure was modified to include only items with adequate discrimination indices. The follow-up ASR condition falsely flagged more children as TD. This demonstrated that ASR sensitivity and specificity did not vary in a single direction. This is a broadly positive finding in that it confirms that ASR is not consistently poorer at identifying a single class of cases. That is, there is no bias toward classifying children as TD or DLD, a positive indicator of ASR's robustness to child-level variations. Furthermore, the mixed results demonstrate that it is possible to modify an ASR item subset to achieve the specific classification metrics necessary for the purposes of the test. For example, an ASR test developed for screening purposes may prefer a higher false positive than false negative rate to ensure that children with DLD are not overlooked. Conversely, an ASR test developed for diagnostic purposes may prefer to have the highest likelihood that a positive result is true to increase confidence that DLD diagnoses are accurate.

A third consideration was that certain item elicitation types may or may not be good candidates for ASR scoring at present. As aforementioned, ASR generally reliably scored test items that required children to complete sentences (i.e., cloze) or that confirmed whether target words or phrases in sentences were repeated (i.e., sentence repetition) and there was no evidence of ASR bias toward either of these item types. However, items requiring children to repeat sentences verbatim yielded inadequate discrimination indices. These findings suggested that the present ASR application was not sensitive enough to reliably confirm whether every word in a given sentence was repeated by a child. Although this study did not compare the word-, phrase-, or sentence-level accuracy of ASR to human transcriptions, this finding is unsurprising given the ASR accuracy rates reported by other researchers. For example, Hair et al. (2019) reported an ASR accuracy rate of 90% on the word-level responses of children with speech disorders. While high, a 90% rate suggests that one in 10 words in a sentence will be incorrectly processed by ASR, indicating that verbatim sentence repetition scoring is likely too stringent for this scoring method at present. It is possible that less stringent criteria for sentence repetition scoring (e.g., 80% of targets detected) may yield higher ASR classification accuracy for this scoring type, but additional analyses are necessary to establish this.

Fourth and finally, this study demonstrated a bias of the ASR tool toward Spanish test items. Specifically, Spanish items achieved substantially higher human–ASR scoring agreement than English items. Additionally, a higher proportion of children in both the TD and DLD groups achieved Spanish best-language scores in the ASR condition relative to the human condition. That is, using ASR transcription and automatic scoring, the diagnostic decision was more often based on children's Spanish test scores than their English test scores. It is likely that, because the children in this study had relatively more exposure to Spanish than English, this pattern of findings did not substantially negatively impact the classification accuracy of the ASR conditions. However, these findings pose a threat to the consideration of both languages of a child, a critical aspect of bilingual language assessment. The results also suggest that characteristics of the English test items made them relatively more challenging for the ASR tool to transcribe. One explanation for this is that English morphosyntactic forms are more likely to be represented by bound morphemes (e.g., kick**ed**) than Spanish morphosyntactic forms, which are more often represented by free morphemes (e.g., **ningun**). Because bound morphemes have lower perceptual salience (Goldschneider & DeKeyser, 2001)—that is, they are shorter in duration, are lower in volume, and/or lack a segment boundary—it is possible that their identification is more challenging for ASR. A more discrete disambiguation of English and Spanish morphosyntactic forms and their features is therefore necessary to further explain the ASR bias toward Spanish test items.

Limitations

There were four primary limitations of this study. First, this study was a secondary analysis of children's existing assessments, which were originally conducted by bilingual examiners. As such, the findings do not establish the practical feasibility of an ASR tool that uses automatic administration and is employed during live dual-language assessment. There is some evidence to suggest that results may differ due to technical challenges arising during live administration (e.g., Yeung et al., 2019). Furthermore, if the current measure is to replace the need for bilingual personnel, it is necessary to establish whether children can complete the tasks without the supervision of

a bilingual practitioner. Therefore, a critical next step for this work is to test automatically administered and scored dual-language tasks during live assessment to determine the extent to which the current results hold and to establish whether it is possible for a practitioner who does not speak the test language to oversee the procedure. Relatedly, the ASR tool used in this study did not handle code-switched responses, such as responses presented in Spanish when English was the test language. This is a behavior that a bilingual practitioner may give credit for (if appropriate) or redirect. Exploring the feasibility of ASR's ability to handle and accept accurate code-switched responses is another necessary step toward ensuring high classification accuracy of the technology for bilingual children.

Second, this study evaluated a small subset of items from a single linguistic domain: morphosyntax. Although test items were culled to maximize classification accuracy, this resulted in an index measure that sampled only one to two exemplars of most morphosyntactic forms. This limited the extent to which conclusions could be made about the morphosyntactic forms, and targets that may be better or worse candidates for ASR assessment. Future work sampling multiple exemplars of each form would provide more evidence for the words and phrases that are best analyzed by ASR, further informing research on automatic expressive language assessment. Finally, looking beyond morphosyntax, children with DLD demonstrate deficits in other language areas (e.g., semantics and narratives) that may also be candidates for ASR assessment and can provide examiners with a broader picture of a child's language skills. Previous work (Albudoor et al., 2019) provided early evidence that a broader subset of test items including probes across linguistic domains could achieve adequate scoring agreement with ASR. However, more research is needed to determine whether a cross-domain ASR measure can yield adequate classification accuracy and/or provide more comprehensive information about a child's language skills.

Third, given that the children in this study did not receive a speech sound production screener or assessment, we were unable to determine whether the ASR tool handled responses containing speech errors differently than those not containing speech errors. This is an important consideration for future work as children with language impairment are more likely to present with speech delays relative to their age-matched peers with TD (Aguilar-Mediavilla et al., 2002). This has implications for the use of ASR for the detection of DLD risk, as an ASR tool that falsely flags responses containing speech errors may overidentify DLD risk among children with typical speech errors or children with speech sound disorders who do not present with co-occurring language deficits.

Fourth, the majority of children in this study spoke one regional dialect of Spanish—Mexican Spanish. This limits the extent to which the findings can be extended to speakers of other regional dialects of Spanish. To ensure that the measure is feasible for use with a broader range of Spanish speakers, additional validation analyses are warranted.

Conclusions

This study provides preliminary support for the technical feasibility of ASR for processing bilingual children's expressive dual-language assessment responses. While the current results are limited in their scope, they represent a proof of concept for the use of ASR in automatic language assessment instruments that test more than one language. Given the barriers to dual-language assessment that have contributed to disproportionate DLD misdiagnosis among bilinguals, this study presents critical evidence toward the expansion of access to more accurate assessment and intervention practices for this population.

Acknowledgments

This research was funded by the National Institute on Deafness and Other Communication Disorders Grant R01 DC010366, awarded to Elizabeth D. Peña.

References

Aguilar-Mediavilla, E. M., Sanz-Torrent, M., & Serra-Raventos, M. (2002). A comparative study of the phonology of preschool children with specific language impairment (SLI), language delay (LD) and normal acquisition. *Clinical Linguistics & Phonetics, 16*(8), 573–596. https://doi.org/10.1080/02699200210148394

Albudoor, N., & Peña, E. D. (2021). Factors influencing US speech and language therapists' use of technology for clinical practice. *International Journal of Language & Communication Disorders, 56*(3), 567–582. https://doi.org/10.1111/1460-6984.12614

Albudoor, N., Peña, E. D., & Bedore, L. M. (2019, November). *Diagnosing language impairment in bilingual children: Speech recognition vs. human scoring* [Presentation]. Irvine Digital Learning Lab, University of California, Irvine, CA.

American Speech-Language-Hearing Association. (2020). *Demographic Profile of ASHA Members Providing Bilingual Services, Year-End 2020*. https://www.asha.org/siteassets/surveys/demographic-profile-bilingual-spanish-service-members.pdf

American Speech-Language-Hearing Association. (2021). *Bilingual service delivery (Practice Portal)*. http://www.asha.org/Practice-Portal/Professional-Issues/Bilingual-Service-Delivery/

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2016). Where Spanish and English come together: A two dimensional bilingual approach to clinical decision making. *Perspectives of the ASHA Special Interest Groups, 1*(14), 3–16. https://doi.org/10.1044/persp1.SIG14.3

Anderson, S. A., Hawes, D. J., & Snow, P. C. (2016). Language impairments among youth offenders: A systematic review.

Children and Youth Services Review, 65, 195–203. https://doi.org/10.1016/j.childyouth.2016.04.004

Arias, G., & Friberg, J. (2017). Bilingual language assessment: Contemporary versus recommended practice in American schools. *Language, Speech, and Hearing Services in Schools, 48*(1), 1–15. https://doi.org/10.1044/2016_LSHSS-15-0090

Audacity Team. (2020). *Audacity* (Version 2.3.2) [Computer program].

Bedore, L. M., Peña, E. D., Anaya, J. B., Nieto, R., Lugo-Neris, M. J., & Baron, A. (2018). Understanding disorder within variation: Production of English grammatical forms by English language learners. *Language, Speech, and Hearing Services in Schools, 49*(2), 277–291. https://doi.org/10.1044/2017_LSHSS-17-0027

Bolt, S. E., & Thurlow, M. L. (2004). Five of the most frequently allowed testing accommodations in state policy. *Remedial and Special Education, 25*(3), 141–152. https://doi.org/10.1177/07419325040250030201

Boster, J. B., & McCarthy, J. W. (2018). Lost in translation: Understanding students' use of social networking and online resources to support early clinical practices. A national survey of graduate speech-language pathology students. *Education and Information Technologies, 23*(1), 321–340. https://doi.org/10.1007/s10639-017-9603-4

Brinson, W., Cook, H., & Wellons, R. (2020). *A systematic review of diagnostic test accuracy for identifying developmental language disorder in bilingual children* [Poster presentation]. Student Research Day, Division of Speech and Hearing Sciences, Department of Allied Health Sciences, University of North Carolina at Chapel Hill. https://doi.org/10.17615/srkb-pc16

Castilla-Earls, A., Bedore, L., Rojas, R., Fabiano-Smith, L., Pruitt-Lord, S., Restrepo, M. A., & Peña, E. (2020). Beyond scores: Using converging evidence to determine speech and language services eligibility for dual language learners. *American Journal of Speech-Language Pathology, 29*(3), 1116–1132. https://doi.org/10.1044/2020_AJSLP-19-00179

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104

de Villiers, J., Iglesias, A., Golinkoff, R., Hirsh-Pasek, K., Wilson, M. S., & Nandakumar, R. (2021). Assessing dual language learners of Spanish and English: Development of the QUILS: ES. *Revista de Logopedia, Foniatría y Audiología, 41*(4), 183–196. https://doi.org/10.1016/j.rlfa.2020.11.001

Dollaghan, C. A. (2007). *The Handbook for Evidence-Based Practice in Communication Disorders*. Brookes.

Dragon Systems, Inc. (1997). *Dragon: Naturally speaking* [Computer program]. SeanSoft.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement*. Prentice-Hall.

Evanini, K., Futagi, Y., & Hauck, M. C. (2020). Using automated scoring in K–12 English language proficiency assessments. In M. K. Wolf (Ed.), *Assessing English language proficiency in U.S. K–12 schools* (pp. 207–225). Routledge.

Gillam, R. B., & Pearson, N. A. (2004). *Test of Narrative Language*. Pro-Ed.

Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Perez, A. (2013). Identification of specific language impairment in bilingual children: I. Assessment in English. *Journal of Speech, Language, and Hearing Research, 56*(6), 1813–1823. https://doi.org/10.1044/1092-4388(2013/12-0056)

Gillam, R. B., Peña, E. D., Bedore, L. M., & Pearson, N. A. (2010). *Test of Narrative Language, Spanish Experimental Version* [Unpublished test].

Goldschneider, J. M., & DeKeyser, R. M. (2001). Explaining the "natural order of L2 morpheme acquisition" in English: A meta-analysis of multiple determinants. *Language Learning, 51*(1), 1–50. https://doi.org/10.1111/1467-9922.00147

Google. (2020). *Google Cloud Speech-to-Text (Version 1) Documentation*. Google Cloud. https://cloud.google.com/speech-to-text/docs/

Hair, A., Ballard, K. J., Ahmed, B., & Gutierrez-Osuna, R. (2019). Evaluating automatic speech recognition for child speech therapy applications. In *The 21st International ACM SIGACCESS Conference on Computers and Accessibility* (pp. 578–580). ACM. https://doi.org/10.1145/3308561.3354606

Hollingshead, A. B. (1975). Four factor index of social status. *Yale Journal of Sociology, 8*, 47–55.

Kim, A. A., Chapman, M., & Banerjee, H. L. (2020). Innovations and challenges in K–12 English language proficiency assessment tasks. In M. K. Wolf (Ed.), *Assessing English language proficiency in U.S. K–12 schools* (Vol. 4, pp. 55–74). Routledge. https://doi.org/10.4324/9780429491689-4

MacArthur, C. A., & Cavalier, A. R. (2004). Dictation and speech recognition technology as test accommodations. *Exceptional Children, 71*(1), 43–58. https://doi.org/10.1177/001440290407100103

Morgan, P. L., Farkas, G., Hillemeier, M. M., Mattison, R., Maczuga, S., Li, H., & Cook, M. (2015). Minorities are disproportionately underrepresented in special education: Longitudinal evidence across five disability conditions. *Educational Researcher, 44*(5), 278–292. https://doi.org/10.3102/0013189X15591157

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., Vamvakas, G., & Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: Evidence from a population study. *Journal of Child Psychology and Psychiatry, 57*(11), 1247–1257. https://doi.org/10.1111/jcpp.12573

Peña, E. D., & Bedore, L. M. (2011). It takes two: Improving assessment accuracy in bilingual children. *The ASHA Leader, 16*(13), 20–22. https://doi.org/10.1044/leader.FTR3.16132011.20

Peña, E. D., Bedore, L. M., & Griffin, Z. (2010). *Cross-language outcomes of typical and atypical development in bilinguals.* National Institute of Deafness and Other Communication Disorders.

Peña, E. D., Bedore, L. M., Iglesias, A., Gutierrez-Clellen, V. F., & Goldstein, B. A. (2012a). *Bilingual English-Spanish Assessment-Middle Elementary (BESA-ME), Experimental Version* [Unpublished test].

Peña, E. D., Bedore, L. M., Iglesias, A., Gutierrez-Clellen, V. F., & Goldstein, B. A. (2012b). *Bilingual English Spanish Oral Screener (BESOS), Experimental Version* [Unpublished test].

Peña, E. D., Bedore, L. M., & Kester, E. S. (2016). Assessment of language impairment in bilingual children using semantic tasks: Two languages classify better than one. *International Journal of Language & Communication Disorders, 51*(2), 192–202. https://doi.org/10.1111/1460-6984.12199

Peña, E. D., Bedore, L. M., Lugo-Neris, M. J., & Albudoor, N. (2020). Identifying developmental language disorder in school age bilinguals: Semantics, grammar, and narratives. *Language Assessment Quarterly, 17*(5), 541–558. https://doi.org/10.1080/15434303.2020.1827258

Peña, E. D., Gutiérrez-Clellen, V. F., Iglesias, A., Goldstein, B., & Bedore, L. M. (2018). *Bilingual English-Spanish Assessment (BESA)*. Brookes.

Pettit, B., & Western, B. (2004). Mass imprisonment and the life course: Race and class inequality in U.S. incarceration. *American Sociological Review, 69*(2), 151–169. https://doi.org/10.1177/000312240406900201

Plante, E., & Vance, R. (1994). Selection of preschool language tests: A data-based approach. *Language, Speech, and Hearing*

Services in Schools, 25(1), 15–24. https://doi.org/10.1044/0161-1461.2501.15

Pratt, A. S., Anaya, J. B., Ramos, M. N., Pham, G., Muñoz, M., Bedore, L. M., & Peña, E. D. (2022). From a distance: Comparison of in-person and virtual assessments with adult-child dyads from linguistically diverse backgrounds. *Language, Speech, and Hearing Services in Schools, 53*(2), 360–375. https://doi.org/10.1044/2021_LSHSS-21-00070

Royal College of Speech and Language Therapists. (2021) *Bilingualism overview*. https://www.rcslt.org/speech-and-language-therapy/clinical-information/bilingualism

Sabu, K., & Rao, P. (2018). Automatic assessment of children's oral reading using speech recognition and prosody modeling. *CSI Transactions on ICT, 6*(2), 221–225. https://doi.org/10.1007/s40012-018-0202-3

Saenz, T. I., & Langdon, H. W. (2019). Speech-language pathologists' collaboration with interpreters: Results of a current survey in California. *Translation & Interpreting, 11*(1), 43–62. https://doi.org/10.12807/ti.111201.2019.a03

Spille, C., Kollmeier, B., & Meyer, B. T. (2018). Comparing human and automatic speech recognition in simple and complex acoustic scenes. *Computer Speech & Language, 52,* 123–140. https://doi.org/10.1016/j.csl.2018.04.003

Sullivan, A. L., & Bal, A. (2013). Disproportionality in special education: Effects of individual and school variables on disability risk. *Exceptional Children, 79*(4), 475–494. https://doi.org/10.1177/001440291307900406

Thompson, S., Blount, A., & Thurlow, M. (2002). *A summary of research on the effects of test accommodations: 1999 through 2001* [Technical report]. National Center on Educational Outcomes, University of Minnesota.

Thurlow, M. L., & Johnson, D. R. (2011). *The high school dropout dilemma and special education students*. University of California, Santa Barbara.

U.S. Census Bureau. (2019). *American Community Survey (ACS)*. https://www.census.gov/programs-surveys/acs

Wood, D. A. (1960). *Test construction: Development and interpretation of achievement tests*. Charles E. Merrill Books, Inc.

Wood, L., Kiperman, S., Esch, R. C., Leroux, A. J., & Truscott, S. D. (2017). Predicting dropout using student-and school-level factors: An ecological perspective. *School Psychology Quarterly, 32*(1), 35–49. https://doi.org/10.1037/spq0000152

Xu, Y., Wang, D., Collins, P., Lee, H., & Warschauer, M. (2021). Same benefits, different communication patterns: Comparing children's reading with a conversational agent vs. a human partner. *Computers & Education, 161,* 104059. https://doi.org/10.1016/j.compedu.2020.104059

Yeung, G., Afshan, A., Quintero, M., Martin, A., Spaulding, S., Park, H. W., Bailey, A., Breazeal, C., & Alwan, A. (2019, April). *Towards the development of personalized learning companion robots for early speech and language assessment*. 2019 Annual Meeting of the American Educational Research Association (AERA), Toronto, Canada. https://par.nsf.gov/biblio/10099084, https://doi.org/10.302/1431402