9-2022

Comparing Measures From Computer-Administered and Examiner-Administered Narrative Retells in Spanish: A Pilot Study

John J. Heilmann
University of Wisconsin - Milwaukee

Denise A. Finneran
University of Oklahoma Health Sciences Center

Maura Jones Moyle
Marquette University, maura.moyle@marquette.edu

American Journal of Speech - Language Pathology, Vol. 31, No. 5 (September 2022): 2175-2185. DOI.

Comparing Measures From Computer-Administered and Examiner-Administered Narrative Retells in Spanish: A Pilot Study

John Heilmann
Department of Communication Sciences and Disorders, University of Wisconsin–Milwaukee
Denise Finneran
Department of Communication Sciences and Disorders, The University of Oklahoma Health Sciences Center, Oklahoma City
Maura Moyle
Department of Speech Pathology and Audiology, Marquette University, Milwaukee, WI

ABSTRACT

Purpose

Narrative language sample analysis (LSA) is a recommended best practice for the assessment of monolingual and bilingual children. With business-as-usual narrative LSA, examiners are actively

involved in all aspects of the elicitation. Software advancements have shown multiple benefits of computer-administered language assessments, some of which may be beneficial for narrative assessments, particularly for bilingual children. The goal of this pilot study was to test the feasibility of computer-administered narrative retells in bilingual children.

Method

Ten English-Spanish bilingual children, kindergarten to fourth grade, completed two narrative retells using wordless picture books (Frog Goes to Dinner and Frog on His Own) in two conditions: examiner-administered and computer-administered. Five narrative measures were generated from these 20 transcripts.

Results

Significant, strong correlations were observed between the two elicitation methods for four of the five measures. We completed a series of Wilcoxon signed-ranks tests and found no significant differences in measures across the elicitation methods. Follow-up descriptive analyses revealed few large differences across elicitation methods for the individual participants.

Conclusion

This study provides preliminary evidence on the use of a computer-administered narrative procedure and motivates further research on the method to confirm its validity and to document its effectiveness within clinical practice.

Supplemental Material: https://doi.org/10.23641/asha.20346648

Clinicians are actively involved in every aspect of the assessment process in business-as-usual speech-language pathology evaluations. For example, common norm-referenced language tests require a speech-language pathologist (SLP) to provide instructions, give prompts, and solicit responses (e.g., Wiig et al., 2013). Common criterion-referenced assessment protocols also require considerable examiner effort, with SLPs providing materials and cues to observe children's performance on language-based tasks (e.g., Invernizzi et al., 2004). Language sample analysis (LSA), another popular assessment practice, also requires substantial examiner involvement when employing business-as-usual practices, with the SLP providing a meaningful communicative context and then engaging with the child to collect a representative sample of their communication (Heilmann, Tucci, et al., 2020). Although examiner-administered assessments are commonplace in speech-language pathology, there are alternatives. Computer-administered assessments employ software to automate many of the assessment processes, such as automated prompts, direct entry of responses by the test taker, and automated scoring. Most states' summative English Language Arts assessments, such as the Wisconsin Forward exam or the California Assessment of Student Performance and Progress, are computer administered, showing that computer-administered language assessments are commonplace for assessing student progress in K-12 general education (National Center for Education Statistics, n.d.). Computer-administered assessments are less commonly used in speech-language pathology clinical practice, but an emerging literature has shown their feasibility within clinical assessment protocols. Upon examining performance on two versions of a norm-referenced vocabulary test completed by 30 adolescents with learning disabilities, Wiig et al. (1996) documented strong correlations between paper-based and computer-administered forms of the test. Carson et al. (2011) found equivalent

performance on a traditional paper-based and computer-administered phonological awareness assessment with a group of 4- and 5-yearold children without disabilities. Polišenská and Kapalková (2014) documented strong reliability and validity on a computer-administered nonword repetition task embedded into a recorded story that was completed by a large sample of 2- to 6-year-old children. These promising results motivate testing of additional computer-administered language assessments.

One assessment that may benefit from computer administration is LSA. Whereas business-as-usual LSA employing standardized procedures can be highly accurate and reliable when implemented by a diverse group of SLPs (Heilmann, Malone, et al., 2020), any examiner-administered assessment can have human error. By automating aspects of the testing process, computer-administered LSA may reduce examiner-level variability that can impact child performance, such as fidelity to protocols and variability in verbal or nonverbal support. For example, Westerveld and Heilmann (2012) showed that narrative retells were significantly less complex when picture cues were withheld from participants, confirming that alterations in elicitation can have a significant impact on language measures. Computer-administered assessments could also improve the efficiency of LSA. Most SLPs serving pediatric populations do not regularly use LSA due to the perceived impracticality (Klatte et al., 2022; Pfeiffer et al., 2019). Computer-administered LSA might present a feasible option that allows SLPs to practice at the "top of the license," focusing on oversight of the comprehensive evaluation and interpretation of the LSA data while potentially delegating the elicitation and other components of the LSA process to support staff (McNeilly, 2018).

Computer automation could assist with a particularly unfeasible aspect of clinical practice: monolingual English-speaking SLPs completing language assessments with children who are not monolingual English speaking. In the most recent Data Snapshot published by the American Speech-Language-Hearing Association (ASHA), only 8% of SLPs were self-identified bilingual providers, meaning that 92% of clinicians do not consider themselves qualified to provide services to clients who speak multiple languages (ASHA, 2021). The language status of the workforce does not meet the needs of the population served, as the most recent American Community Survey revealed that 22% of those over the age of 5 years speak a language other than English (U.S. Census Bureau, 2019). To complete an accurate assessment, the examiner must document proficiency of all languages spoken by the client (ASHA, 2004). SLPs working in school settings who are compliant with the Individuals with Disabilities Education Act (IDEA, 2004) must administer assessments in whatever language(s) yield accurate information on the student's academic, developmental, and functional abilities. The guidance is clear-SLPs need to be able to assess in multiple languages-yet the high percentage of monolingual English speakers in the current SLP workforce impedes the ability to adequately serve a linguistically diverse population. The stakes are high, as traditional assessment practices often lead to incorrect diagnostic decisions for bilingual children (Rose et al., 2022) and can lead to some children being disproportionately overidentified and others disproportionately underidentified (Yamasaki & Luk, 2018).

Narrative LSA has multiple features that make it a strong candidate for computer administration, which may be particularly useful for bilingual children. Compared to common standardized tests and more open-ended language sampling tasks, which require frequent examiner instructions and prompts throughout the testing session (e.g., Wiig et al., 2013), narrative retells are less interactive and entail

greater independence in the children's performance. In the elicitation of a narrative retell, the examiner typically provides instructions and then reads a scripted story to the child (with or without a picture book). The child is then given a set of instructions and completes a narrative production, often with minimal guidance from the examiner (e.g., Gagarina et al., 2012; Hiebert & Rojas, 2021). When the examiner is not required to make real-time decisions about prompts and feedback, examiner input could potentially be automated through computerized administration. Once the sample is elicited, the recorded sample could be outsourced to a trained transcriber who has proficiency in the child's language. There is also a strong clinical rationale for choosing narratives when considering the type of assessments to computerize for bilingual children. Narrative assessment tasks are, in general, culturally familiar to children from a variety of backgrounds (Gagarina et al., 2012). Multiple studies have documented the power of narrative measures in predicting reading and academic outcomes (Huang et al., 2022; Miller et al., 2006), making narrative LSA a recommended tool for use within the comprehensive assessment of bilingual children (Castilla-Earls et al., 2020; Ebert & Pham, 2017).

There have been many successful examples of computer-administered language assessments in K-12 general education (National Center for Education Statistics, n.d.) and an emerging literature showing the potential for computer-administered language assessments in clinical speech-language pathology (Carson et al., 2011; Polišenská & Kapalková, 2014; Wiig et al., 1996). However, there is no published systematic investigation on Spanish LSA for computer-administered as compared to examiner-administered story retells. In this study, we completed pilot testing of a computer-administered narrative retell procedure to address this gap in the literature. If children's performance on computer-administered narrative retells is equivalent to business-as-usual examiner-administered narrative retells, we anticipated that measures from these two elicitation methods would be similar. On the other hand, we know that differences in elicitation method can have a significant impact on narrative measures. For example, Westerveld and Heilmann (2012) found that children produced longer, more complex stories when they had pictures to support their narrative retells and speculated that the presence of pictures reduced the cognitive load, which assisted with the recall of story components. Diehm et al. (2020) compared narratives elicited using animated videos and static pictures and found that the more engaging video-based elicitation method resulted in longer, more complex narrative productions. We may expect similar types of elicitation effects in this study. If a certain elicitation method is more cognitively taxing, children may produce narratives with less complex language. If the presence of a live examiner is more engaging than interacting with a computer, children may produce longer, more linguistically complex narratives. If a certain elicitation method is more efficient, children may produce shorter, more fluent narratives. To test these predictions, we collected computer- and examiner-administered narrative retells from a pilot sample of English-Spanish bilingual children.

Our first aim was to determine if children's performance on the computer-administered narratives aligned with their performance on examiner-administered narratives. To address this aim, we completed correlation analyses that rank-ordered the children within each elicitation method and then compared the rank orderings across the two elicitation methods. This analysis allowed us to determine if the children's scores were related across the two conditions but provided limited information on groupwise differences across conditions. For example, if every child scored exactly 50% higher on one of the elicitation measures, their scores would be perfectly correlated but there would be a sizeable group difference. Therefore, our second aim was to determine if most children's narrative measures

had larger values when the sample was collected using a particular elicitation method. If the children favored the examiner-administered narratives, we expected that most children would score higher on their examiner-administered narratives. Our final aim was to explore differences in measures across elicitation methods for each individual child. We sought to explain instances where individual children's performance varied drastically across elicitation methods. We interpreted these different pilot data to determine if further study of computer-administered narratives is warranted. To make that determination, we addressed the following research questions.

1. Are there strong positive correlations between measures of sample length, verbal fluency, and language complexity that were generated from computer-administered narratives when compared to the same measures generated from examiner-administered narratives?
2. Do most Spanish-English bilingual children produce longer, more fluent, and more linguistically complex narratives when using computer-administered or examiner-administered narrative elicitation methods?
3. Can we easily explain the largest individual differences in narrative measures across elicitation methods?

Method

Participants

Prior to initiating the study, the University of Wisconsin-Milwaukee institutional review board approved all proposed project activities. Recruitment and data collection were completed by a research assistant (RA) majoring in communication sciences and disorders under the direction of the first author. The RA was bilingual with native proficiency in both Spanish and English. We recruited 10 English-Spanish bilingual children with good Spanish proficiency, grades kindergarten through fourth, to participate in this pilot study. Six of the children were recruited through two local elementary schools, and the remaining four children were recruited through word-of-mouth advertising. Each participant's parent provided signed informed consent prior to participating in the study. Two participants were in kindergarten, five were in second grade, one was in third grade, and three were in fourth grade. There were an equal number of boys and girls. For the six participants from the elementary schools, school personnel confirmed that each student was English-Spanish bilingual and was receiving specialized instruction designed for English language learners. For all participants, an adult who knew the child well confirmed that the child had good Spanish proficiency and answered "yes" to the question, "Can {child's name} have a basic conversation with someone in Spanish?" In addition, school personnel or parents confirmed that each child identified as Latinx and was not receiving special education services.

Narrative Retell Protocol

For both elicitation methods, children listened to a scripted story while following along with the pictures from a wordless picture book. The children were then presented with the first picture in the book and instructed to retell the story using the pictures to support their production. For this study, we used two of the Frog series of wordless picture books by Mercer Mayer, which have been used widely in speech-language pathology research (e.g., Hiebert & Rojas, 2021; Squires et al., 2014) and are integrated into the narrative retell protocols of the Systematic Analysis of Language Transcripts (SALT;

Miller & Iglesias, 2020). Heilmann et al. (2016) examined Spanish-English bilingual children's performance on five of the Frog wordless picture books and found that the measures from each story were generally equivalent. In this study, we used two stories that generated very similar narrative measures in Spanish (Frog on His Own and Frog Goes to Dinner; Heilmann et al., 2016). The presentation of the stories was evenly counterbalanced across elicitation methods. The order of elicitation methods was also evenly counterbalanced across the participants.

Due to the ongoing global pandemic, we used a videoconferencing application for all sessions, which had become a standard practice for clinical practice throughout the pandemic (Campbell & Goldstein, 2021). Even though there are no data showing equivalence of measures from narrative retells for elementary school-age children when comparing performance in person and using a videoconferencing application, we suspected that there would be modest differences given the similarity in measures across settings for story retells in adults with brain injuries (Brennan et al., 2004) and for toddlers completing play-based language samples with their parents (Manning et al., 2020).

Examiner-Administered Narrative Protocol

With the examiner-administered elicitation method, the examiner described the task to the children in Spanish while they viewed the examiner's face on their screens. After the children confirmed that they understood the task, the examiner shared her computer screen, which contained pictures from the respective Frog picture book, which were scanned and embedded into a PowerPoint file. The main screen displayed the book's pictures, and an embedded screen showed the examiner's face. The examiner read the scripted story in Spanish while sharing each respective picture from the book. After completing the presentation, the examiner asked the children to retell the story in Spanish. During the retell, the children had access to the pictures and the examiner provided minimal cueing except when needed for encouragement (e.g., "Keep going" or "You're doing great!"). During the retell, the examiner advanced the pictures when it was clear that the child completed each respective segment of the narrative. If the children did not provide clear cues, the examiner silently counted for 5 s after they finished a sentence. If the children provided no additional information, she advanced to the next page.

Computer-Administered Narrative Protocol

We purchased an annual subscription for the Online Story Elicitation Program from SALT Software, LLC. For the computer-administered elicitation method, the examiner told each child to listen and follow the directions. The examiner then started the elicitation program, which began with a cartoon avatar on the screen describing the task in Spanish. After providing instructions, the recorded speaker narrated the Spanish story script while the corresponding pictures, which were embedded into the software, were automatically presented on the screen. The avatar then returned and provided instructions for the children's production of the story in Spanish. The children saw the pictures and produced their own narrative production without any input from the avatar or examiner. During the child's narrative production, the examiner used the same page-advancing procedure described above.

Transcription

All language samples were recorded using Audacity and transcribed using SALT (Miller & Iglesias, 2020), which is specialized software to assist with transcription and analysis of language samples across multiple languages. SALT has specialized features to assist with Spanish transcription, including the ability to transcribe diacritic characters, routines to identify Spanish root words, and specialized

coding for reflexive and nonreflexive pronouns. The utterances were segmented using conversation units (C-units) modified for the features of Spanish narration. Traditional C-units include an independent clause and all associated dependent clauses. When children tell stories in Spanish, many speakers will produce strings of clauses that do not contain a pronoun that would be obligatory when narrating in English (Rojas & Iglesias, 2006). Utterances with chains of verbs with no subjects were segmented using the modified C-unit segmentation rule. See the studies of Miller et al. (2019) and Miller and Iglesias (2020) for a full description of the transcription rules.

Because this was a pilot study and there were only 10 participants, our priority was using a process that would result in highly accurate transcription. The RA completed an initial transcription of each sample. Prior to transcribing the samples in this study, she completed 10 hr of transcription training using online modules (SALT Software, n.d.). She transcribed two practice samples that were checked by a Spanish-fluent speech-language pathology student with over 50 hr of experience as a paid transcriber. After correcting her errors and refining her transcription skills, the RA transcribed the 20 language samples from the 10 participants. These transcripts were sent to a third Spanish-fluent transcriber, who had hundreds of hours of experience as a professional transcriber of children's Spanish language samples. The third transcriber was blind to the purpose of the project and to the elicitation method for each sample. This transcriber listened to the sample (starting after the instructions and with the child's first produced word) and made corrections to the original transcript.

Narrative Measures

We used SALT to generate five microstructural measures commonly reported in studies of Spanish-speaking children's narrative skills (Hiebert & Rojas, 2021; Miller et al., 2019; Squires et al., 2014). We also considered narrative measures that may be susceptible to differences across elicitation methods. Total words was based on the total number of words produced by the children in the sample, excluding those that were unintelligible or were a reduplication or reformulation (i.e., mazes). If one of the elicitation methods was more engaging or if the expectations were clearer to the participants, we anticipated that they may produce more words in their samples. Elapsed time was based on the number of minutes from the start to the finish of the children's narrative productions. If a particular elicitation method was less efficient, we anticipated an increase in the time it took to produce the sample. Words per minute was used as a measure of verbal fluency, calculated by dividing all words produced in the sample (including those that were unintelligible, in mazes, and in abandoned utterances) by the elapsed time. If an elicitation method resulted in a disjointed narrative production, we anticipated a decrease in verbal fluency. The final two measures were included to determine if the elicitation method had an impact on measures of language ability. Mean length of C-unit (MLC-u) in words is a popular and widely used measure of utterance length in Spanish speakers (Rojas & Iglesias, 2006). It was calculated in this study by summing the total words and dividing by the number of modified C-units. Number of different words (NDW), an index of lexical diversity, was calculated by summing the total number of different root words produced in the sample. If a particular elicitation context were more taxing and took greater cognitive resources, we anticipated a decrease in these two language-based measures. For a full description of the SALT measures, see Miller et al. (2019) or Miller and Iglesias (2020).

Results

Table 1 summarizes the descriptive statistics for each measure across the two elicitation methods. On average, the narratives were roughly 300 words in length (range: 170-420 words) and took 6 min (range: 3.9-7.6 min) to for the children to produce. As observed in the table, mean values were similar across the two conditions. A series of nonparametric Spearman's rho correlations were calculated for each measure, comparing measures from the computer- and examiner-administered narrative tasks. As observed in the bottom row of Table 1, there were significant positive correlations for four of the five measures. We further explored the impact of elicitation method on MLC-u because it was the only measure that had a nonsignificant correlation. Upon examining the descriptive statistics and individual children's performance, we noticed that most of the children, coincidentally, had similar MLC-u values. To document the amount of variance across the measures, we calculated a single coefficient of variation (CV) for each measure by dividing the standard deviation by the mean. The CV provides a summary of the group variance relative to the mean, providing equivalence in comparing variance across the measures. This analysis confirmed that there was more limited variance in MLC-u (CV = .08) when compared to total words, elapsed time, WPM, and NDW (CV = .21, .18, .23, and .16, respectively), which may explain, in part, the weaker correlation observed for MLC-u relative to the other measures.

Table 1. Descriptive statistics (M and SD) and correlations for measures from computer- and examiner-administered narrative retells

Condition	Total words	Elapsed time	Words per minute	MLC-u	NDW
Computer-administered	306.0 (72.7)	6.0 (1.2)	62.2 (14.4)	7.7 (0.8)	91.7 (18.0)
Examiner-administered	302.4 (60.1)	5.9 (1.0)	62.7 (15.7)	7.8 (0.6)	89.7 (11.4)
R	.88*	.77*	.81*	.58	.85*

Note. MLC-u = mean length of C-unit; number of different words = NDW.
*$p \leq .01$.

To determine if a particular elicitation method had a higher likelihood of generating larger values for each measure, we completed a series of Wilcoxon signed-ranks tests. We chose a nonparametric test because of the small sample size. Through the signed-ranks tests, we calculated the difference across the two elicitation methods for each measure (e.g., computer-administered total words – examiner-administered total words). Then, we ranked the absolute values of the differences in measures, with the smallest difference ranked 1 and the largest difference ranked 10. The ranked values for the positive and negative differences were separately summed, providing two numbers that totaled 55 (i.e., 1+2 + 3 + 4 + 5 + 6 + 7 + 8 + 9+ 10). The final W-statistic was the smaller of the two summed rankings. A W-statistic based on negative differences (i.e., W-) would indicate that the computer-administered narratives generated larger values on measures for most children, and a W-statistic based on positive differences (i.e., W+) would indicate that the examiner-administered narratives generated larger values on measures for most children. A W-statistic close to zero would indicate that most of differences (and especially the large differences) between conditions went the same direction. For example, a W-statistic of 1 would indicate that computer-administered elicitation generated larger values for one participant (and that difference was the smallest difference between conditions for all participants) and that examiner-administered elicitation generated higher values for the other nine

participants. Conversely, a larger W-statistic would indicate that the direction of the differences between elicitation methods was more evenly split across the participants (and/or, there were no differences for participants). For example, a W-statistic close to 27.5 (i.e., 55 = 2) would indicate that half of the children did better on the examiner-administered narratives and the other half of the children did better on the computer-administered narratives.

Based on our sample size of n =10, Wilcoxon and Wilcox (1964) provided the two-tailed critical value of W = 8 for significance at p = .05. In this study, our observed W-statistics were greater than eight for each measure (W = 19-25), indicating that there were no statistically significant differences for any of the measures. We used SPSS to convert the W-statistics to z scores and to generate more precise p values. All z scores were < 1.0, and all p values were .39 or greater, showing continued nonsignificance when using more liberal critical values for each measure (total words [W+ = 25, z = 0.26, p = .80], elapsed time [W+ = 19, z = 0.87, p = .39], words per minute [W- = 24, z = 0.36, p = .72], MLC-u [W- = 19, z = 0.87, p = .39], and NDW [W+ = 22.5, z = 0.51, p = .61]). In summary, results from the Wilcoxon signed-ranks tests showed that the direction of differences across the conditions did not follow a regular pattern for any of the measures.

Our final exploratory analysis described the change in measures across elicitation methods for each of the 10 participants. We calculated percent change scores by subtracting the examiner-administered measure from the computer-administered measure, then dividing by the computer-administered measure. Positive values were generated when the computer-administered measures were larger, and negative values were generated when the examiner-administered measures were larger. To identify the children whose narrative measures changed the most across elicitation methods, we computed the average percent change by calculating the mean of the absolute values of percent change for each individual measure (summarized in Table 2).

Table 2. Percent change from computer-administered to examiner-administered elicitation methods.

Participant	Total words	Elapsed time	Words per minute	MLC-u	NDW	Average percent change
1	.13	.09	.01	−.12	.18	.11
2	.10	.12	.15	−.04	.05	.09
3	.02	−.07	−.13	−.01	−.15	.08
4	.00	−.02	.01	−.05	.05	.03
5	.02	.14	−.08	.02	−.04	.06
6	−.39	−.42	.03	−.04	−.14	.20
7	−.09	.07	−.05	−.05	−.07	.07
8	−.02	−.02	−.12	−.06	.09	.06
9	−.05	.01	−.04	.09	−.04	.05
10	.18	.08	.17	.07	.12	.12

Note. Average percent change values calculated using absolute values of percent change for each individual measure. MLC-u = mean length of C-unit; NDW = number of different words.

Most of the individual changes were modest, with less than 10% change across conditions for most of the measures from most of the participants. Of the 60 percent change calculations, 42 were less than 10%. To illustrate some of the more pronounced changes across elicitation methods, we describe the

narrative productions of the three participants with the largest average changes across conditions (Participants 1, 6, and 10). Some of the bigger changes in measures across elicitation contexts involved the length of the samples, reflected in total words, elapsed time, and, to an extent, NDW. Participant 6 produced a longer sample during examiner elicitation and Participants 1 and 10 produced longer samples during computer elicitation. A particular story or the order of elicitation could be responsible for the differences in measures, but a review of Table 3 shows that there was good counterbalancing across the stories, elicitation methods, and order of presentation, making story or order effects unlikely.

Table 3. Elicitation protocols for three participants with the largest average changes in measures across elicitation methods.

Participant		Computer-administered	Examiner-administered
1	Story	Frog on His Own	Frog Goes to Dinner
	Order	2	1
6	Story	Frog Goes to Dinner	Frog on His Own
	Order	2	1
10	Story	Frog Goes to Dinner	Frog on His Own
	Order	1	2

We reviewed transcripts to explore potential explanations for the changes across conditions. Participant 6, who had some of the largest observed changes in measures across conditions, was interrupted by school announcements multiple times during the examiner-administered retell, resulting in multiple extended pauses that added to the total time required to produce the sample. The pauses may also explain why the participant produced more total and different words-having forced pause times may have provided an advantage for formulating thoughts, recalling portions of the narrative script, and producing more detail. Conversely, Participants 1 and 10 produced notably shorter samples with examiner administration despite no interruptions. To explore additional explanations for differences across conditions, we used SALT to complete a lexical field analysis on multiple word classes (conjunctions, modals, multiple types of pronouns, and determiners; see Supplemental Material S1). Upon examining use of definite and indefinite article (el, la, los, las, un, uno, unos, and unas), we found that the longer, computer-administered samples from Participants 1 and 10 contained more articles (Participant 1: computer-administered = 54 articles, examiner-administered = 45 articles; Participant 10: computer-administered = 40 articles, examiner-administered = 22 articles). These articles were attached to the different characters and objects in the story, meaning that longer samples contained more detailed descriptions of the characters and plot in their computer-administered samples. Participant 6 produced slightly more articles in the longer, examiner-administered narratives, but the difference was not as striking as Participants 1 and 10 (examiner-administered = 41, computer administered = 36). Participant 6 did use more coordinating conjunctions in the examiner-administered narrative (n = 38) when compared to computer-administered narratives (n = 23), meaning that the child may have produced a longer narrative by separating story components into more independent clauses.

A final cross-condition observation was a different pattern in Participant 1's and 6's use of modal auxiliaries, which are used to add concepts of time and mood to a sentence. Participant 1 had longer

MLC-u in the examiner-administered narratives and produced notably more modal auxiliaries in the examiner-administered narrative (n = 15) when compared to computer-administered narratives (n = 9). Facchinetti et al. (2003) described how modal auxiliary use can be tied to the use of dependent clauses, which could explain a possible relationship with longer utterances. Participant 6 also demonstrated more frequent use of modal auxiliaries in the examiner-administered narratives (n = 13) when compared to computer-administered narratives (n = 3), but with no associated changes in MLC-u across conditions.

Discussion

The goal of this study was to provide preliminary data on Spanish measures generated from computer-administered narrative retells collected from bilingual children. We achieved this goal by comparing performance on computer-administered narrative retells to performance on business-as-usual examiner-administered narrative retells. Upon examining the properties of five narrative measures, we documented strong correlations between the two elicitation methods for most of the measures. The one exception to this trend was a nonsignificant correlation between the elicitation methods for MLC-u. Our subsequent analysis showed that the children in our sample produced narratives with similar MLC-u values, which may have impacted the results of the correlation analysis. The selection of Spearman's rho correlations, which are nonparametric, was motivated by the small sample size. This correlation approach rank-ordered the individuals within each elicitation method and compared the rankings across elicitation methods. With limited variability in overall performance as demonstrated by the relatively small CV for MLC-u, small differences across the elicitation methods had a notable impact on a child's relative ranking. We designed this study to prevent homogeneity in performance by sampling from a wide range of grades (kindergarten through fourth grade), but the participants unexpectedly had similar MLC-u values. We excluded children who were receiving speech/language services, yet many bilingual Latinx children are misidentified (Yamasaki & Luk, 2018); hence, there is a chance that some children had unidentified language disorders. Another possibility is that there may have been variability in Spanish proficiency when compared to English proficiency. For example, the younger children may have had relatively stronger Spanish proficiency than the older children. Because we were interested in testing the consistency across conditions, we only required children to have a basic level of Spanish proficiency and did not measure relative proficiency across English and Spanish. We believed that leaving relative proficiency free to vary was consistent with clinical practice, where examiners test children with differing levels of proficiency across languages spoken. Even though there were nonsignificant correlations between the elicitation methods for MLC-u, the subsequent analyses showed that children were equally likely to have a higher MLC-u in either elicitation method, suggesting that the children, as a group, performed similarly across the two elicitation methods.

The goal of the Wilcoxon signed-ranks tests was to determine if a particular elicitation method was associated with larger values on measures for most of the participants. If the children, as a group, found one of the elicitation methods more awkward, taxing, or tedious, we expected that the affiliated narrative measures would have lower values when compared to measures from the more engaging, less taxing elicitation method. We did not find any such systematic patterns across elicitation methods. Because elicitation method did not systematically impact the length of the produced samples, the children's verbal fluency, and the measures of general language ability, we believe that the children

had a similar experience across both elicitation methods and that the elicitation method had little impact on the children's ability to demonstrate their narrative language skills.

Our final analysis, which examined changes in measures across elicitation methods for the individual participants, provided further evidence that most differences across elicitation methods were modest, with less than 10% change in most measures across conditions. The two most striking individual differences across conditions, elapsed time and total words for Participant 6, could be explained by idiosyncratic interruptions during that particular session. In clinical practice, SLPs would need to interpret these data with a high level of caution due to the variation from standard protocol. Our further exploration of the three children who had the largest average differences across conditions showed no easy explanation for the observed differences. Our confirmation of counterbalancing for these participants verified that these differences were not an artifact of the story used or order of elicitation. The lexical field analysis suggested that Participant 1's and 10's longer, computer-administered narratives could be related to the increased use of articles tied to descriptions of objects and characters, yet article use had a more modest association with Participant 6's longer sample. The observed differences in Participant 1 's use of modal auxiliaries could explain the cross-condition differences in MLC-u, but frequent modal use across conditions was not related to MLC-u for Participant 6. We concluded that our lexical field analysis did not assist with identifying consistent explanations for the differences across conditions. Because most of the cross-condition differences were small and the small number of larger differences were not easily explained, we believe that individual differences across elicitation methods were likely due to some unexplained factor, such as measurement error, rather than systematic differences due to the elicitation method.

Our results extend prior research showing equivalence in computer- and examiner-administered narratives in adults with brain injury (Brennan et al., 2004) and in toddlers completing play-based samples with their parents (Manning et al., 2020). This replicated pattern shows that computer-administered language assessments may be feasible across a range of ages and language tasks. Given the well-documented clinical utility of business-as-usual narrative retells (Castilla-Earls et al., 2020; Ebert & Pham, 2017; Gagarina et al., 2012; Rojas & Iglesias, 2006), these preliminary data suggest that examiners have options for how to elicit their narrative samples, as dictated by the needs of the examiner and client.

Clinical Implications

The evidence from this pilot study provides emerging guidance for clinical practice. Clinicians interested in eliciting narrative retells themselves in the language(s) they speak may find modest benefits from computer-administered narrative procedures. Computer-administered narrative language sampling provides a high level of standardization that promotes consistency. Nonetheless, SLPs following a standardized protocol can also elicit samples with a high level of fidelity (Heilmann, Malone, et al., 2020). In addition, computer administration offers modest time savings, as the elicitation of a narrative retell is inherently a quick process. In this study, it took children 6 min, on average, to produce their narratives across both conditions. If an examiner feels that there are clinical advantages to eliciting computer-administered narratives, these preliminary data show that the resulting data will likely be equivalent to examiner-administered narrative retells.

Other applications of computer-administered narrative retells may have a more meaningful impact on clinical practice. If some of the specialized elicitation procedures could be completed by the computer, SLPs may be able to enlist the assistance of related personnel, such as SLP assistants or teacher aides, to assist with administration. This approach aligns with the top of the license approach to clinical practice, where SLPs limit their practice to tasks that require their unique skillset and cannot be delegated to nonclinicians (McNeilly, 2018). If certain language assessments, such as narrative retells, can be elicited in an automated way and transcribed without direct involvement of the highly skilled SLP, those less specialized tasks could be delegated to support staff. SLPs could then focus their attention on the highly specialized skill of interpreting children's performance within the comprehensive assessment. This same top of the license approach could apply when assessing children who speak languages not spoken by the examiner. Because there are considerably more bilingual clients than bilingual SLPs available to serve them (ASHA, 2021; U.S. Census Bureau, 2019), SLPs need tools to obtain valid data in languages they do not speak. The first priority is to increase the cultural and linguistic diversity of the SLP workforce to better meet the needs of the population (Fannin & Mandulak, 2021). Until then, computer-administered narratives may help SLPs acquire some assessment data across languages. Because narrative retells do not require substantial real-time interactive guidance from an examiner, all that is needed are basic instructions and a script in the child's language. Our results suggest that presenting an audio recording with a computerized avatar may be sufficient for eliciting a representative sample of a child's narrative skills. As with any assessment, examiners should ensure that the task and format are familiar to the child and that they will provide meaningful clinical data.

Limitations and Future Directions

Because this was a pilot study with a small sample size, these data should be interpreted with caution when used to motivate clinical or research practices. In addition to a small sample size, there were further limitations that should be considered when interpreting the data. We did not measure Spanish proficiency, which would have better described the characteristics of our participants and could have helped explain the homogeneity in MLC-u across such a wide range of grades. Of the dozens of measures that can be generated from narrative language samples, we limited our focus to five microstructural measures that we thought would be sensitive to differences in elicitation method. Further study would be required to determine the impact of computer administration on additional narrative measures, including macrostructural aspects of overall narrative quality. Another limitation of this study was that business-as-usual practice was not so usual. This study was completed during a global pandemic, where universities, schools, and families were trying to minimize unnecessary exposure to COVID-19. We completed both elicitation contexts through a screen to optimize safety of the children and examiner. Comparing in-person examiner-administered narrative retells to computer-administered retells may have been more authentic and more likely to reveal differences across the elicitation methods, but it simply was not a possibility. As noted by Campbell and Goldstein (2021), videoconferencing had become a standard practice during the pandemic and will likely continue to be part of some clinical practices moving into the future.

These data motivate further study of computer-administered language assessments across language tasks and beyond the Spanish language. Our pilot data showing general equivalence across the two

elicitation methods are an important first step. Additional research testing on computer-administered narrative retells within clinical contexts is needed to fully validate the methodology. SLPs also need an evidence base showing the value and suitability for multiple stakeholders, including parents, teachers, other professionals, and, of course, the clients themselves.

Clinical language assessments are complex, requiring quality tools and clinical expertise for administration and interpretation. Our results show the promise of computer-administered narrative retells, particularly for bilingual children. This work contributes one small piece to the comprehensive assessment puzzle. When assessing children's language skills, SLPs must use converging sources of evidence, collected across all languages spoken, to make accurate determinations about levels of functioning (Castilla-Earls et al., 2020; Ebert & Pham, 2017). Further research and development of tools to assist with these comprehensive assessments, including computer-administrated language assessments, will provide the evidence base necessary for this multifaceted process.

Acknowledgments

Funding for the participant incentives was provided to the first author by the College of Health Sciences at the University of Wisconsin-Milwaukee. Additional funding for this project was provided to the first author by the Support for Undergraduate Research Fellows at the University of Wisconsin-Milwaukee. The authors would like to thank Tania Ortega, Katrina Reeder, and Sarah Sterk for their assistance with this project.

Sidebar

Correspondence to John Heilmann: heilmanj@uwm.edu. Disclosure: The authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

References

American Speech-Language-Hearing Association. (2004). Preferred practice patterns for the profession of speech-language pathology [Preferred Practice Patterns]. https://0-www-asha-org.libus.csd.mu.edu/policy/ PP2004-00191/#sec1.3.19

American Speech-Language-Hearing Association. (2021). Demographic profile of ASHA members providing bilingual services. https://0-www-asha-org.libus.csd.mu.edu/siteassets/surveys/demographic-profilebilingual-spanish-service-members.pdf

Brennan, D. M., Georgeadis, A. C., Baron, C. R., & Barker, L. M. (2004). The effect of videoconference-based telerehabilitation on story retelling performance by brain-injured subjects and its implications for remote speech-language therapy. Telemedicine Journal and E-Health, 10(2), 147-154. https:// doi.org/10.1089/tmj.2004.10.147

Campbell, D. R., & Goldstein, H. (2021). Genesis of a new generation of telepractitioners: The COVID-19 pandemic and pediatric speech-language pathology services. American Journal of Speech-Language Pathology, 30(5), 2143-2154. https://0-doi-org.libus.csd.mu.edu/ 10.1044/2021_AJSLP-21-00013

Carson, K., Gillon, G., & Boustead, T. (2011). Computer-administrated versus paper-based assessment of school-entry phonological awareness ability. Asia Pacific Journal of Speech, Language and Hearing, 14(2), 85-101. https://0-doi-org.libus.csd.mu.edu/10.1179/ 136132811805334876

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://0-doi-org.libus.csd.mu.edu/10.1044/2020_AJSLP-19-00179

Diehm, E. A., Wood, C., Puhlman, J., & Callendar, M. (2020). Young children's narrative retell in response to static and animated stories. International Journal of Language & Communication Disorders, 55(3), 359-372. https://0-doi-org.libus.csd.mu.edu/10.1111/14606984.12523

Ebert, K. D., & Pham, G. (2017). Synthesizing information from language samples and standardized tests in school-age bilingual assessment. Language, Speech, and Hearing Services in Schools, 48(1), 42-55. https://0-doi-org.libus.csd.mu.edu/10.1044/2016_LSHSS-160007

Facchinetti, R., Krug, M. G., & Palmer, F. R. (Eds.). (2003). Modality in contemporary English (Vol. 44). Walter de Gruyter. https://0-doi-org.libus.csd.mu.edu/10.1515/9783110895339.xv

Fannin, D. K., & Mandulak, K. C. (2021). Introduction to the forum: Increasing diversity in the communication sciences and disorders workforce, part 1. American Journal of SpeechLanguage Pathology, 30(5), 1913-1915. https://0-doi-org.libus.csd.mu.edu/10. 1044/2021_AJSLP-21-00258

Gagarina, N. V., Klop, D., Kunnari, S., Tantele, K., Välimaa, T., Balčiūnienė, I., Bohnacker, U., & Walters, J. (2012). MAIN: Multilingual Assessment Instrument for Narratives. ZAS Papers in Linguistics, 56, Article 155. https://0-doi-org.libus.csd.mu.edu/10.21248/ zaspil.56.2019.414

Heilmann, J., Malone, T. O., & Westerveld, M. F. (2020). Properties of spoken persuasive language samples from typically developing adolescents. Language, Speech, and Hearing Services in Schools, 51(2), 441-456. https://0-doi-org.libus.csd.mu.edu/10.1044/2019_ LSHSS-19-00078

Heilmann, J., Rojas, R., Iglesias, A., & Miller, J. F. (2016). Clinical impact of wordless picture storybooks on bilingual narrative language production: A comparison of the 'Frog' stories. International Journal of Language & Communication Disorders, 51(3), 339-345. https://0-doi-org.libus.csd.mu.edu/10.1111/1460-6984.12201

Heilmann, J., Tucci, A., Plante, E., & Miller, J. F. (2020). Assessing functional language in school-aged children using language sample analysis. Perspectives of the ASHA Special Interest Groups, 5(3), 622-636. https://0-doi-org.libus.csd.mu.edu/10.1044/2020_ PERSP-19-00079

Hiebert, L., & Rojas, R. (2021). A longitudinal study of Spanish language growth and loss in young Spanish-English bilingual children. Journal of Communication Disorders, 92, 106110. https://0-doi-org.libus.csd.mu.edu/10.1016/j.jcomdis.2021.106110

Huang, B. H., Bedore, L. M., Ramírez, R., & Wicha, N. (2022). Contributions of oral narrative skills to English reading in Spanish-English Latino/a dual language learners. Journal of Speech, Language, and Hearing Research, 65(2), 653-671. https://0-doi-org.libus.csd.mu.edu/10.1044/2021_JSLHR-21-00105

Individuals With Disabilities Education Act, 20 U.S.C. 1400. (2004). https://sites.ed.gov/idea/

Invernizzi, M., Sullivan, A., Meier, J., & Swank, L. (2004). Phonological awareness literacy screening: Pre-kindergarten. University of Virginia. https://pals.virginia.edu/public/pdfs/rd/tech/PreK_ technical_chapter.pdf

Klatte, I. S., van Heugten, V., Zwitserlood, R., & Gerrits, E. (2022). Language sample analysis in clinical practice: Speechlanguage pathologists' barriers, facilitators, and needs. Language, Speech, and Hearing Services in Schools, 53(1), 1-16. https://0-doi-org.libus.csd.mu.edu/10.1044/2021_lshss-21-00026

Manning, B. L., Harpole, A., Harriott, E. M., Postolowicz, K., & Norton, E. S. (2020). Taking language samples home: Feasibility, reliability, and validity of child language samples conducted remotely with video chat versus in-person. Journal of Speech, Language, and Hearing Research, 63(12), 3982-3990. https://0-doi-org.libus.csd.mu.edu/10.1044/2020_JSLHR-20-00202

McNeilly, L. (2018). Why we need to practice at the top of the license. The ASHA Leader, 23(2), 10-11. https://0-doi-org.libus.csd.mu.edu/10. 1044/leader.FMP.23022018.10

Miller, J. F., Andriacchi, K. D., & Nockerts, A. (2019). Assessing language production using SALT Software: A clinician's guide to language sample analysis (3rd ed.). SALT Software, LLC. https://www.saltsoftware.com/products/referencebook/saltreference-book-3rd-edition

Miller, J. F., Heilmann, J., Nockerts, A., Iglesias, A., Fabiano, L., & Francis, D. J. (2006). Oral language and reading in bilingual children. Learning Disabilities Research & Practice, 21(1), 30-43. https://0-doi-org.libus.csd.mu.edu/10.1111/j.1540-5826.2006.00205.x

Miller, J. F., & Iglesias, A. (2020). Systematic Analysis of Language Transcripts (Version 20) [Computer software].

National Center for Education Statistics. (n.d.). State education practices (SEP). https://nces.ed.gov/programs/statereform/tab2_ 22.asp

Pfeiffer, D. L., Pavelko, S. L., Hahs-Vaughn, D. L., & Dudding, C. C. (2019). A national survey of speech-language pathologists' engagement in interprofessional collaborative practice in schools: Identifying predictive factors and barriers to implementation. Language, Speech, and Hearing Services in Schools, 50(4), 639-655. https://0-doi-org.libus.csd.mu.edu/10.1044/2019_LSHSS-18-0100

Polišenská, K., & Kapalková, S. (2014). Improving child compliance on a computer-administered nonword repetition task. Journal of Speech, Language, and Hearing Research, 57(3), 1060-1068. https://0-doi-org.libus.csd.mu.edu/10.1044/1092-4388(2013/13-0014)

Rojas, R., & Iglesias, A. (2006). Bilingual (Spanish-English) narrative language analyses: Why and how? SIG 14 Perspectives on Communication Disorders and Sciences in Culturally and Linguistically Diverse (CLD) Populations, 13(1), 3-8. https:// doi.org/10.1044/cds13.1.3

Rose, K., Armon-Lotem, S., & Altman, C. (2022). Profiling bilingual children: Using monolingual assessment to inform diagnosis. Language, Speech, and Hearing Services in Schools, 53(2), 494-510. https://0-doi-org.libus.csd.mu.edu/10.1044/2021_lshss-21-00099

SALT Software. (n.d.). Self-paced online training. https://www. saltsoftware.com/training/self-paced-online-training

Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Giliam, R. B. (2014). Story retelling by bilingual children with language impairments and typically developing controls. International Journal of Language & Communication Disorders, 49(1), 60-74. https://0-doi-org.libus.csd.mu.edu/10.1111/14606984.12044

U.S. Census Bureau. (2019). Selected social characteristics in the United States [Data file]. https://www.census.gov/acs/www/ data/data-tables-and-tools/data-profiles/

Westerveld, M. F., & Heilmann, J. J. (2012). The effects of geographic location and picture support on children's story retelling performance. Asia Pacific Journal of Speech, Language and Hearing, 15(2), 129-143. https://0-doi-org.libus.csd.mu.edu/10.1179/jslh.2012.15.2.129

Wiig, E. H., Jones, S. S., & Wiig, E. D. (1996). Computer-based assessment of word knowledge in teens with learning disabilities. Language, Speech, and Hearing Services in Schools, 27(1), 21-28. https://0-doi-org.libus.csd.mu.edu/10.1044/0161-1461.2701. 21

Wiig, E. H., Semel, E., & Secord, W. (2013). Clinical Evaluation of Language Fundamentals-Fifth Edition (CELF-5). Pearson. https://www.pearsonassessments.com/store/usassessments/en/Store/ProfessionalAssessments/Speech-%26-Language/Clinical-Evaluation-ofLanguage-Fundamentals-%7C-Fifth-Edition/p/100000705.html

Wilcoxon, F., & Wilcox, R. A. (1964). Some rapid approximate statistical procedures. Lederle Laboratories.

Yamasaki, B. L., & Luk, G. (2018). Eligibility for special education in elementary school: The role of diverse language experiences. Language, Speech, and Hearing Services in Schools, 49(4), 889-901. https://0-doi-org.libus.csd.mu.edu/10.1044/2018_LSHSS-DYSLC18-0006