JSLHR

Research Article

# Dynamic Assessment of Narrative Ability in English Accurately Identifies Language Impairment in English Language Learners

Elizabeth D. Peña,[a] Ronald B. Gillam,[b] and Lisa M. Bedore[a]

**Purpose:** To assess the identification accuracy of dynamic assessment (DA) of narrative ability in English for children learning English as a 2nd language.
**Method:** A DA task was administered to 54 children: 18 Spanish–English-speaking children with language impairment (LI); 18 age-, sex-, IQ- and language experience-matched typical control children; and an additional 18 age- and language experience-matched comparison children. A variety of quantitative and qualitative measures were collected in the pretest phase, the mediation phase, and the posttest phase of the study. Exploratory discriminant analysis was used to determine the set of measures that best differentiated among this group of children with and without LI.
**Results:** A combination of examiner ratings of modifiability (compliance, metacognition, and task orientation), DA story scores (setting, dialogue, and complexity of vocabulary), and ungrammaticality (derived from the posttest narrative sample) classified children with 80.6% to 97.2% accuracy.
**Conclusion:** DA conducted in English provides a systematic means for measuring learning processes and learning outcomes, resulting in a clinically useful procedure for identifying LIs in bilingual children who are in the process of learning English as a second language.

When children are in the process of learning a second language it is difficult to know whether low performance on a language measure is due to lack of language experience or to compromised language learning ability. Accurate assessment of bilingual children is a critical practical need in the field (Bedore & Peña, 2008; Dollaghan & Horner, 2011; Kohnert, 2010; Thordardottir, Rothenberg, Rivard, & Naves, 2006). In addition to current work identifying markers of language impairment (LI) for different languages (e.g., Leonard, 2014) there is a need to develop methods that may help clinicians identify LI in bilingual individuals' second language (Gillam, Peña, Bedore, Bohman, & Mendez-Pérez, 2013; Paradis, Schneider, & Duncan, 2013).

*Dynamic assessment* (DA) has been proposed as a strategy for assessing language in children from culturally and linguistically diverse populations (Carlson, 1983; Carlson & Wiedl, 1980; Hasson & Joffe, 2007; Laing & Kamhi, 2003). A mediated learning experience (MLE) approach to DA is thought to minimize assessment bias related to lack of experience (Carlson & Wiedl, 1980; Laing & Kamhi, 2003) because the focus is on measuring the learning process rather than measuring static knowledge, which is subject to cultural and linguistic bias. In the present study we extended DA procedures to examine a group of children who were in the process of learning English as a second language. These children were highly variable in their language performance due to individual differences in experience with each language (Kohnert, 2010; Thordardottir, 2010) and in language learning abilities (Hadley, Rispoli, Fitzgerald, & Bahnsen, 2011; Hayiou-Thomas, Dale, & Plomin, 2012). This variation makes it difficult to accurately diagnose LI. Because DA focuses on learning rather than norm comparisons, we explored whether it can provide clinically useful information that can guide diagnostic decisions when it is applied in children's second language.

## *DA of Language Ability*

Our approach to DA includes clinician observation of the cognitive and affective strategies children use while actively involved in a language learning task such as narrative. This focus on learning strategies is consistent with an information-processing perspective of LI. In this view, children with LI are proposed to have difficulties learning language due primarily to inefficiencies in attention and memory (Ellis Weismer, Evans, & Hesketh, 1999; Ellis

[a]University of Texas at Austin
[b]Utah State University, Logan
Correspondence to Elizabeth D. Peña: lizp@mail.utexas.edu

Editor: Rhea Paul
Associate Editor: Margarita Kaushanskaya
Received June 11, 2013
Revision received November 15, 2013
Accepted July 10, 2014
DOI: 10.1044/2014_JSLHR-L-13-0151

***Disclosure:*** *The authors have declared that no competing interests existed at the time of publication.*

Weismer et al., 2000; Gillam, Montgomery, & Gillam, 2009; Kohnert, Windsor, & Yim, 2006). DA is designed to support and evaluate children's use of learning strategies such as attention, memory, and cognitive flexibility during teaching as an index of a child's capacity for learning.

DA includes three phases—test, teaching, and retest—creating a context for assessing a child's modifiability. During the first testing phase, the examiner collects data relevant to a particular language domain. During the teaching phase, the examiner provides MLEs using teaching strategies that are designed to increase task performance (Kozulin, 2002; Lidz, 2002; Tzuriel, 2000) and rates the child's responsiveness to instruction. These ratings focus on observations of the cognitive and affective strategies children use during mediated learning. In the retest phase, which is usually conducted within 2 weeks of the teaching phase, language measures are repeated. Children's modifiability is measured by comparing their performance on the pretest and posttest behavioral measures and/or assessing the examiner's qualitative ratings of their responsiveness to instruction. Children who make large behavioral gains and/or efficiently use cognitive strategies (e.g., task orientation, attention, flexibility, and problem solving) are posited to have normal language learning abilities (Burton & Watkins, 2007; Camilleri & Law, 2007). In contrast, small behavioral gains and/or difficulty in efficiently deploying strategies during the teaching phase are thought to be associated with LI.

DA using a test–teach–retest approach has been most commonly applied to the assessment of children from culturally and linguistically diverse backgrounds (Gutiérrez-Clellen & Peña, 2001; Laing & Kamhi, 2003). A number of studies have demonstrated DA's utility for evaluation of language tasks such as word learning (Peña, Iglesias, & Lidz, 2001; Peña, Quinn, & Iglesias, 1992), narrative production (Kramer, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Ukrainetz, Harpell, Walsh, & Coyle, 2000). In most of these studies, typically achieving children made greater gains than children with LIs after intervention. However, there has been some inconsistency with respect to whether such gains alone result in better classification accuracy. In contrast, examiner observations of child modifiability during the teaching phase has consistently differentiated children with LI (or risk for LI) and those with typical language ability with acceptable levels of diagnostic accuracy. For example, Peña et al. (2006) found that measures of modifiability and posttest scores together accurately distinguished children with and without LI with 100% accuracy. In a follow-up analysis of the individual indicators of modifiability, Peña, Reséndiz, and Gillam (2007) found that observations of metacognition and flexibility together classified children with and without LI with 93% accuracy.

### *DA With English Language Learners*

Two recent studies provide support for using DA in English to assess the language skills of English language learner children. Kapantzoglou, Restrepo, and Thompson (2012) gave a novel word learning task to 4- and 5-year-old Spanish speakers learning English. Bilingual Spanish–English speaking children were taught sets of real words and nonsense words that corresponded to Spanish phonotactics. The words were taught in Spanish using a mediated learning script that was presented in three teaching sessions. Probes conducted after each teaching session showed that children with typical development learned new words with fewer exposures. Observations of modifiability using a learning strategies checklist yielded significant differences between children with and without LI. A combination of word identification and modifiability best discriminated among children with and without LI. The results of Kapantzoglou et al.'s study indicates that it is possible to examine child modifiability as they learn both real and nonsense words. Thus, it may be possible to observe strategy use in children with and without LI even in a language with which they have less experience.

Hasson, Camilleri, Jones, Smith, and Dodd (2013) assessed the diagnostic accuracy of the Dynamic Assessment of Preschoolers' Proficiency in Learning English. The intervention phase involved structured cueing of phonological, vocabulary, and sentence structure targets in English, which was the children's second language. Bilingual children between ages 3 and 5 who were receiving speech-language therapy were compared to a control group of typically achieving, age-matched bilingual children. The typically achieving controls made greater pretest–posttest gains than children who were receiving language therapy. Hasson et al.'s study demonstrates that children with and without risk for LI may exhibit differential pretest–posttest performance in their second language.

In the current study, we evaluated the DA of English narration using the dynamic assessment and intervention (DAI) narrative learning task described by L. Miller, Gillam, and Peña (2001). Although narrative content may differ across cultures in terms of expectations (Minami, 2008; Minami & McCabe, 1995), purpose (Nicolopoulou, 2002), and structure (Wang & Leichtman, 2000), narration is a common discourse context across cultures. Thus, narrative assessment may result in lower levels of test bias than standardized testing. For example, Cleave, Girolametto, Chen, and Johnson (2010) found that monolingual English and dominant English bilingual children with LIs produced stories of similar length and complexity even though the bilingual children scored lower on standardized measures of language (e.g., Clinical Evaluation of Language Fundamentals–Preschool:2, Wiig, Secord, & Semel, 2005; and Structured Photographic Expressive Language Test, Werner & Kresheck, 1983). Another advantage of narratives is that they can be elicited by a wordless picture book, which constrains the task in ways that facilitate comparisons across individuals, groups, and time. Finally, teaching a discourse-level task enables investigators to examine the cognitive and social-emotional strategies children use as they construct stories in the teaching phase of DA.

Our four research questions included the following:

1. What are the patterns of narrative learning from pretest to posttest based on language ability?

2. Are there differences in strategy use by language ability as indicated by observation of modifiability?
3. What combination of story and modifiability measures best differentiate children by language ability in matched and comparison samples?
4. Does the diagnostic accuracy of DA differ between children without impairment who were closely matched to the LI children and children without impairment who were not closely matched?

## Method

### *Participants*

Participants were 18 children with LI; 18 children with normal language development matched on age, sex, language experience, and IQ (the *NL-Match group*), and an additional 18 children with normal language development matched only on age and language experience (the *NL-Compare group*) for a total of 54 children. The participants were drawn from a sample of 167 children who participated in a 2-year longitudinal study of diagnostic markers of LI (Gillam et al., 2013). A large group of 1,198 preschoolers were screened in English and Spanish prior to entering kindergarten. A subset of 167 children who used English and Spanish at least 20% of the time and scored at or below the 30th percentile on the Bilingual English Spanish Oral Screener (BESOS; Peña, Bedore, Gutiérrez-Clellen, Iglesias, & Goldstein, 2008) on at least one Spanish subtest and one English subtest were invited to the study. As a group, the Bilingual English Spanish Oral Screener test scores for this subset of 167 children were within the range of scores earned by the larger group of bilingual respondents from which they were drawn. However, this approach increased the likelihood that we would recruit children with LI as well as children whose profiles might lead to a misdiagnosis of LI.

Of the 167 children in the longitudinal portion of the study, 21 children were identified with primary LI on the basis of independent ratings of three speech-language pathologists with expertise in bilingualism. The remaining 146 children were available for selection as matches.

To create a normal language control group (NL-Match), each of the children identified with LI was matched to a typically developing child based on sex, age in months at time of initial testing (within 5 months; *M* difference = 1.86), month of birth (within 4 months; *M* difference = 1.31 months), IQ (within 15 points, *M* difference = 9.05), score on the Universal Nonverbal Intelligence Test (Bracken & McCallum, 1998), and language experience. These are the same matches as reported by Squires et al. (2014). Language experience ratings included percentage of language input and output in Spanish and English and age at first English exposure according to parent and teacher report (Bohman, Bedore, Peña, Mendez-Perez, & Gillam, 2010; Gutiérrez-Clellen & Kreiter, 2003; Restrepo, 1998). Percentage of language input and output was averaged over that reported at pre-kindergarten, kindergarten, and first grade. Matches were selected to be within 20% English and Spanish input and output (average match was within 7.58%). Year at first English exposure data was also averaged across parent report collected at the three testing time points. At the individual level, each matched pair was within 0.85 year. There were no significant differences among the means on any of the matching variables (*p*s >.05). Children's language ability ratings were 4.09 (*SD* = 0.43) for the NL-Match group.

A second comparison control group (NL-Compare) was created for the purpose of cross-validating the findings. Cross-validation allows evaluation of how the discriminant model may perform under less constrained, more realistic conditions. We matched a second NL child to each of the 18 children with LI using age at first English exposure, age in months, and percentage of English and Spanish input and output using the same criteria as above. To improve the generalization of the cross-validation findings, we did not match on IQ or sex. Average language ability ratings for the NL-Compare group was 3.65 (*SD* = 0.54).

For the three groups combined, average exposure to Spanish and English in kindergarten was 44.13% and 55.87%, respectively. Note that at the individual level children ranged from 15% to 92% exposure to English. IQ scores averaged 94.89 on the Universal Nonverbal Intelligence Test. Parent educational and occupational levels were similar across the three groups. This information is displayed in Table 1 by group for descriptive purposes.

### *Procedure*

Because there are no agreed-on, validated gold standard measures for identifying LI in bilingual individuals, we adopted the procedures used by Records and Tomblin (1994) and Tomblin, Records, and Zhang (1996) to establish the diagnosis of LI. Following Tomblin and his colleagues, expert biliingual clinicians with a history of providing services to bilingual children rated the children's language ability on the basis of their clinical expertise (also see Tomblin, 2006). Three experienced bilingual clinicians provided an independent rating of children's language abilities based on a review of the range of formal and informal language data collected when children were in first grade. The kindergarten data were not used because we wanted to give children in the study opportunity for stabilization of first and second language performance (Baker, 2001). Rating all the children data collected during first grade allowed raters to calibrate their ratings to the entire range of performance after the kindergarten year, thereby increasing within-rater consistency.

Raters independently reviewed each of the 167 children's test protocols and transcribed narrative samples, as well as parent and teacher responses to questionnaires concerning Spanish and English language history, current language usage, and proficiency judgments collected during the first-grade year. All three raters were blind to any previous diagnoses, and they did not have access to the children's standard scores for any of the measures. For each case, one of the three expert raters had knowledge of the child either through direct or indirect observation or testing. The other

**Table 1.** Demographic data.

| Characteristic | LI (*n* = 18) | NL-Match (*n* = 18) | NL-Compare (*n* = 18) | ANOVA | | LSD post hoc |
|---|---|---|---|---|---|---|
| | | | | *F* | *p* | |
| Age in months | 68.44 (4.84) | 68.78 (3.81) | 69.78 (4.35) | 0.457 | .636 | |
| Mother occupation | 1.88 (2.09) | 1.83 (2.23) | 1.38 (2.03) | 0.286 | .753 | |
| Father occupation | 2.94 (1.39) | 3.17 (1.58) | 2.44 (1.41) | 1.082 | .347 | |
| Mother education | 2.47 (1.70) | 3.50 (1.82) | 2.69 (1.74) | 1.674 | .198 | |
| Father education | 2.53 (1.97) | 3.33 (2.06) | 2.69 (1.70) | 0.860 | .430 | |
| % Kindergarten English | 55.98% (20.41%) | 56.62% (22.26%) | 55.00% (20.37%) | 0.027 | .973 | |
| % Kindergarten Spanish | 44.02% (20.41%) | 43.38% (22.26%) | 45.00% (20.37%) | 0.027 | .973 | |
| Age at first English exposure | 2.13 (1.33) | 2.04 (1.39) | 1.83 (1.43) | 0.201 | .819 | |
| UNIT | 88.72 (12.01) | 93.61 (13.12) | 102.33 (10.87) | 5.906 | .005 | NL-Compare > LI*** |

*Note.* Numbers in parentheses are standard deviations. Education and occupation scores were adapted from Hollingshead (1975). Educational scores range from 0 (*no formal schooling*) to 7 (*graduate professional training*). Mean scores from 2 to 3 are consistent with a ninth- to 11th-grade education. Occupational scores range from 0 (*unemployed*) to 9 (*major professionals*). Mean scores between 1 and 3 are consistent with menial skill workers, unskilled workers, and semiskilled workers. LI = children with language impairment; NL-Match = children with normal language development matched on age, sex, language experience, and IQ; NL-Compare = children with normal language development matched only on age and language experience; ANOVA = analysis of variance; UNIT = Universal Nonverbal Intelligence Test; LSD = least significant difference, no post hoc adjustments.

***$p$ = .001.

two expert raters had no previous contact with the child or the child's records.

The intent was for the raters to use their knowledge of LI profiles of bilingual children to determine whether patterns they observed in the data were expected or unexpected for a given age. Consistent with Tomblin et al.'s (1996; Records & Tomblin, 1994) clinical judgment procedure, our expert judges rated three domains (narration, vocabulary–semantics, and grammar), and scored each domain on a 6-point scale (0 = severe/profound impairment, 1 = moderate LI, 2 = mild impairment, 3 = low normal performance, 4 = normal performance, and 5 = above normal performance). Ratings were completed for each language. After consideration of the three domains, the raters were asked to make overall summary ratings for each language. A child was viewed as having an LI when the summary scores of at least two of the three expert raters were 2, 1, or 0 (indicating mild, moderate, or severe LI, respectively) in the child's best language. Average ratings were LI = 1.74 ($SD$ = 0.63), NL = 4.09 ($SD$ = 0.44), and NL-Compare = 3.72 ($SD$ = 0.57). Across the 167 children, the overall point-to-point agreement among the three raters was 90%.

### *DA*

DA was conducted during the children's kindergarten year using the procedures outlined in DAI (L. Miller et al., 2001) and Peña et al. (2006). Children told stories in English from the wordless picture book *Two Friends* (L. Miller, 2000b) as the pretest story and *Bird and His Ring* (L. Miller, 2000a) as the posttest story. The DA was conducted over three sessions over a 7- to 14- day period. The first session included the pretest and the first intervention, the second session included the second intervention and the third session included the posttest. All examiners were blind to child ability during testing and DAI.

*MLE sessions.* Two MLE sessions, which were each 30 min long, were conducted at least 2 days apart in English. Examiners followed the procedures outlined by L. Miller et al. (2001) and Peña et al. (2006). The scripted interventions (see online Appendix) focused on increasing the length and complexity of the children's narratives. Before implementation of the interventions, we reviewed the scripts to identify words or phrases that might be misunderstood by kindergarten-age children who were in the process of learning English. The content and structure of the intervention stayed consistent with that described by Peña et al., with the exception that we simplified all grammatically complex sentences (e.g., "When people tell stories they include a number of parts" was changed to "Complete stories have lots of parts").

The MLE sessions focused on modeling and practicing the creation of complete and complex episodes. Macrostructure aspects of stories, such as inclusion of character information, temporal markers, and causal relationships, were emphasized to a lesser extent in the teaching scripts. During the first session the clinician reviewed the story the child had told during the pretest and worked with the child to make the story more complete and complex. During the second session the clinician and the child coconstructed a story for the wordless picture book *A Boy, a Dog, a Frog, and a Friend* (Mayer & Mayer, 1971). The online Appendix contains the complete scripts for both sessions.

The story intervention scripts incorporated key instructional elements of MLE (Lidz, 1991, 2002), such as intention to teach, meaning, transcendence, planning, and transfer. *Intention to teach* focused on explaining the learning goal to the child, (e.g., "Today, we're going to talk about telling complete stories"). *Mediation of meaning* demonstrated that the goal was important (e.g., "It's important to be able to tell good stories"). *Transcendence* helped the child relate the goal to everyday activities (e.g., "You read and write

stories at school"). *Planning* encouraged the child to think about the overall goals (e.g., "Tell me what the important parts of a story are again?"). Finally, *transfer* encouraged the child to use the strategies they were taught (e.g., "How are you going to remember to include all these parts of the story?").

*Mediated learning observation*. At the end of the first session, examiners rated child responsivity using the mediated learning observation (MLO) form available in Peña et al.'s (2007) article. Only one session was rated because a previous study of MLO showed there was a very high level of consistency in scoring of mediated learning across sessions ($r > .95$; Peña et al., 2007). The MLO form consisted of 12 items. There were three items for each of four areas: (a) affect (anxiety, motivation, persistence), (b) behavior (responsiveness to feedback, attention, compliance), (c) arousal (task orientation, metacognition, nonverbal self-reward), and (d) elaboration (e.g., problem solving, flexibility, verbal mediation). Each item was rated from 1 (*requiring little examiner support*) to 5 (*requiring maximum examiner support*), with lower scores representing better responsiveness to mediation.

*Clinician training*. Seven different clinicians conducted the mediated learning sessions. All were bilingual Spanish–English speakers with clinical and/or research experience with young children. Five of the clinicians were speech-language pathologists certified by the American Speech-Language-Hearing Association; the other two were bachelor's-level research associates with extensive background working with bilingual children in school and research settings. Clinicians were trained to use the intervention scripts and to rate child modifiability using the mediated learning observation form (Peña et al., 2007). Clinicians watched videotaped examples of MLE sessions while comparing them with the scripts. The Mediated Learning Experience Rating Scale (Lidz, 1991) was used to guide observation and discussion of implementation of the MLE sessions. Training on the MLO form was conducted in a similar manner. After discussion of the observation form, scoring procedures and examples, clinicians watched videotapes of previous sessions and rated child modifiability. Their scores were compared to the original scores and discussed.

Two of the clinicians had participated in previous studies of DA and had implemented MLE for narratives. They conducted the first MLE sessions, which were videotaped. The rest of the clinicians observed the sessions, scored the observation form, and discussed their ratings. Once clinicians were familiar with the scripts and materials and were able to reliably score child modifiability, they implemented their first MLE sessions. These were videotaped and reviewed to ensure consistency in following the scripts.

*Intervention fidelity*. Videotapes of 26 MLE sessions were randomly selected for evaluation of fidelity. The Mediated Learning Experience Rating Scale was used to rate the level at which clinicians included each of the critical components of MLE (i.e., intention to teach, meaning, transcendence, planning, and transfer). Each of these components was rated by an independent scorer not associated with the study using a 4-point scale, ranging from 0 (*not included*) to 3 (*consistently included with a statement of principle given by the examiner*). An average rating was 2.91 across the 26 sampled sessions, with no ratings below 2. This indicated consistent implementation of MLE.

## Narrative Measures

Pretest and posttest narratives were collected as children created stories that were related to the two wordless picture books from the DAI. The books were balanced with respect to number of characters, episodes, and attempts and have been found to yield similar scores on measures of story components, story ideas, episode structure, and productivity (Peña et al., 2006). Before and after intervention, children were asked to look through all the pictures and scenes in a book and to think of a story to tell based on the pictures. After looking through all the pictures together, the examiner prompted the child to return to the beginning of the book and to tell a story. Examiners used minimal cues during examination of the book; during elicitation of the story they used only back-channeling cues to prompt the child to continue.

Children's stories were audiorecorded with digital recorders and were transcribed using Systematic Analysis of Language Transcripts (SALT; J. Miller & Iglesias, 2008) transcription conventions for word and sentence segmentation. In addition, each utterance was coded for grammaticality (grammatical/ungrammatical) and complexity (number of main verbs). Stories were scored on 10 qualitative items that yield a total DAI story score. Items were divided into three subsections: (a) story components (setting, character information, temporal order of events, and causal relationships), (b) story ideas and language (complexity of ideas, knowledge of dialogue, complexity of vocabulary, grammatical complexity, and creativity), and (c) episode structure (initiating event, internal response, attempt, reaction, and resolution). Each item of the story components and ideas-and-language category scores were rated using a 5-point scale and summed within category (e.g., story components, ideas and language). The episode structure score was based on a 7-point scale that categorized combinations of elements included in the story. The total DAI story score was the sum of all category scores plus the episode structure score and ranged from 10 to 52. A higher score was representative of a better quality narrative.

We also calculated basic SALT-derived measures for productivity, including total number of words (TNW), number of different words (NDW), and mean length of utterance in words (MLUw); a measure of grammaticality (percentage ungrammatical utterances); and a measure of complexity (number of main verbs).

## Scoring and Reliability

Story transcripts were first checked for accuracy by having a second person listen to the audiofiles while reviewing the transcription. Once two transcribers agreed on the

words and the sentence segmentation, the transcripts were coded for story components (setting: time and place, character information, and temporal order of events), ideas and language (grammatical complexity, complexity of idea, creativity) and episode structure. A second coder independently scored 32 pairs of pretest and posttest stories (64 stories total) that were randomly selected from the 167 pairs of stories that had been coded. Point-to-point agreement for the individual story item scores was 87%, and agreement within 1 point was 98%. The Pearson product–moment correlation of .91 for story items as scored by the initial coder and the reliability coder demonstrated a high level of agreement.

## Results

### *Pretest and Posttest Comparisons by Group*

The first set of analyses addressed potential changes in narrative ability in response to the two MLE sessions. The dependent measure was the total DAI story score at pretest and posttest (see Table 2). We calculated a mixed analysis of variance (ANOVA) with group (LI, NL-Match, and NL-Compare) as the between-subjects factor and time (pretest and posttest) as the within-subject factor. For the total DAI story score, Mauchly's tests of sphericity yielded an alpha level less than .05, indicating that we did not meet sphericity assumptions. Therefore, we report the multivariate Wilks's λ results, which do not require sphericity because they are based on difference scores that are computed by comparing scores from each level of the within-subject factor. There was a significant main effect of group, $F(2, 51) = 3.608$, $p = .034$, $\eta_p^2 = .124$; and time, $F(1, 51) = 4.722$, $p = .034$, $\eta_p^2 = .085$. Pairwise comparisons indicated that children in the LI group scored significantly lower ($M = 16.22$) than the NL-Match group ($M = 20.36$, $p = .030$) and the NL-Compare ($M = 20.69$, $p = .020$) group. There were no differences between the two NL groups ($p = .858$). Across groups, children scored higher on the posttest story ($M = 19.70$) compared to the pretest story ($M = 18.48$). The Group × Time interaction was not significant.

To assess potential group differences in productivity and complexity, we computed a series of mixed ANOVAs to examine MLUw, NDW, TNW, percentage ungrammaticality, and number of main verbs per utterance. Group was the between-subjects factor, and time was the within-subject factor. Again, we report the multivariate solution because assumptions of sphericity were not met. To minimize the false discovery rate, a Benjamini and Hochberg (1985) adjustment was used to estimate an adjusted significant level for the comparisons. This procedure controls the rate of false discovery by ordering $p$ values from smallest to largest. The smallest $p$ value is multiplied by the number of observations to set the adjusted $p$ value. The next smallest $p$ value is multiplied by the number of observations and divided by its rank. The adjusted $p$ value is compared to that of the next in rank and the smaller of the two is retained. We used a false discovery rate of .05 and here report the adjusted $p$ values.

Results for NDW, TNW, MLUw, and number of main verbs indicated no significant main effects of time or group and no significant Group × Time interactions. For proportion of ungrammatical utterances, there was a significant main effect of group, $F(2, 51) = 6.544$, $p = .045$, $\eta_p^2 = .204$. Neither the time main effect nor the Group × Time interaction reached significance. Children with LI produced a greater proportion of ungrammatical utterances (.63) compared to the NL-Match ($M = .40$) and NL-Compare ($M = .42$) groups. The mediation sessions may have affected narrative performance, but they did not appear to have much of an effect on language productivity, which was not directly addressed in intervention.

### *Observation of Modifiability*

Children's responsivity to mediation was observed during intervention and was rated by the examiners immediately after completion of the first session using the MLO protocol. Recall that the MLO concerns examiner judgments of child learning performance in the areas of affect, arousal, elaboration, and behavior. Ratings for each of these components included three scores ranging from 1 to 5, with higher scores representing greater amounts of clinician support. We summed the three scores to create component scores with

**Table 2.** Pretest–posttest comparisons: Story scores, productivity, and complexity by group.

| | LI | | NL-Match | | NL-Compare | |
|---|---|---|---|---|---|---|
| **Measure** | **Pretest** | **Posttest** | **Pretest** | **Posttest** | **Pretest** | **Posttest** |
| DAI total story score | 15.94 (5.20) | 16.50 (5.83) | 20.11 (5.93) | 20.61 (5.38) | 19.39 (6.20) | 22.00 (6.26) |
| Productivity measures | | | | | | |
| MLUw | 4.80 (1.43) | 4.42 (1.57) | 4.82 (1.18) | 5.08 (1.01) | 4.98 (0.94) | 5.25 (0.92) |
| NDW | 25.28 (14.02) | 25.28 (16.94) | 30.33 (12.65) | 31.83 (11.35) | 32.50 (9.82) | 31.83 (10.03) |
| TNW | 54.00 (32.96) | 64.50 (45.27) | 65.28 (28.29) | 75.22 (35.13) | 76.22 (31.73) | 76.78 (27.75) |
| Complexity measures | | | | | | |
| Number of main verbs | 11.56 (8.05) | 13.61 (11.30) | 15.00 (7.28) | 17.06 (8.98) | 17.67 (7.87) | 15.89 (5.37) |
| Proportion of ungrammatical utterances | .58 (.31) | .68 (.24) | .41 (.29) | .39 (.24) | .37 (.16) | .46 (.22) |

*Note*. Numbers in parentheses are standard deviations. MLUw = Mean length utterance in words; NDW = number of different words; TNW = total number of words.

a possible range of 1 to 15. Note that lower scores are better because they indicate that less examiner support or redirection was needed during the intervention.

We conducted a mixed ANOVA to compare potential group differences on the four subscores of the MLO. Group (LI, NL-Match, NL-Compare) was the between-subjects factor and MLO subscore (affect, arousal, elaboration, behavior) was the within-subject factor. Mauchly's tests of sphericity indicated an alpha level less than .05 so, as before, we report the multivariate solution. Results indicated a main effect of group $F(2, 51) = 9.262$, $p < .001$, $\eta_p^2 = .266$, and of MLO subscore, $F(3, 49) = 25.536$, $p < .001$, $\eta_p^2 = .610$. Pairwise comparisons indicated lower (better) scores for the NL-Match ($M = 5.597$) and NL-Control ($M = 5.597$) groups compared to the LI group ($M = 7.569$, $ps < .001$). There were no significant differences between the NL-Match and NL-Control groups ($p = 1.00$). Pairwise comparisons of MLO subscores demonstrated that children scored better on behavior ($M = 5.074$) and affect ($M = 5.370$) than arousal ($M = 7.278$) and elaboration ($M = 7.296$, $ps < .001$). Behavior and affect scores were not significantly different from each other ($p = .888$). The arousal and elaboration scores were also not significantly different ($p = .137$). The Group × MLO subscore interaction was not significant.

To further understand the contribution of each of the items on the MLO relative to previous work with monolingual children, we compared the three groups on each of the 12 item scores using ANOVA. Results are displayed in Table 3. As before, we used a Benjamini and Hochberg (1985) adjustment to reduce the false discovery rate. The adjusted $p$ values are reported in the table. In general, children with LI required more support during intervention than their typically developing counterparts. There were no differences between children with and without LI on measures of anxiety, nonverbal self-reward, verbal mediation, and responsiveness to feedback.

## Classification Analysis

The third research question concerned the discriminant accuracy of the static and dynamic measures. We were particularly interested in what set of measures best classified bilingual children with and without LI. We conducted the analysis in two steps. First, a reduced set of possible predictors for the discriminant analysis was derived through multiple regression. These predictors were then tested using discriminant analysis.

*Multiple regression*. Three multiple-regression models were explored to address the question of which combination of DA measures converge with language ability. Measures included individual items from the MLO (12 Likert scale scores), individual items from the posttest story scores (10 Likert scores), and language sample SALT-derived scores (five measures). In this analysis, the goal was to determine a parsimonious model that explained the most variance in ability and contained the fewest number of independent variables. Three selection methods—forward selection, backward selection, and stepwise selection—were compared to the full model in order to select the best variable set for the discriminant analysis. The zero-order correlations and summary statistics from the four models, with language ability as the dependent variable, are displayed in

**Table 3.** Modifiability scores by group.

| Item | Group | | | Univariate ANOVA | | Post hoc | $p^a$ |
|---|---|---|---|---|---|---|---|
| | LI | NL-Match | NL-Compare | $F$ | $p^a$ | | |
| Anxiety | 1.61 (0.50) | 1.44 (0.51) | 1.33 (0.49) | 1.410 | .317 | | *ns* |
| Motivation | 2.44 (0.71) | 1.78 (0.65) | 1.83 (0.79) | 4.821 | .046 | LI > NL-Match | .016 |
| | | | | | | LI > NL-Compare | .025 |
| Nonverbal persistence | 2.28 (0.75) | 1.83 (0.92) | 1.56 (0.62) | 3.987 | .025 | LI > NL-Compare | .016 |
| Task orientation | 3.06 (0.94) | 2.11 (0.83) | 2.28 (1.02) | 5.262 | .046 | LI > NL-Match | .013 |
| | | | | | | LI > NL-Compare | .027 |
| Metacognition | 3.50 (0.86) | 2.33 (0.84) | 2.39 (0.98) | 9.751 | .004 | LI > NL-Match | .005 |
| | | | | | | LI > NL-Compare | .005 |
| Nonverbal self-reward | 2.39 (1.04) | 1.89 (0.58) | 1.89 (0.68) | 2.403 | .168 | | *ns* |
| Problem solving | 2.94 (0.87) | 2.00 (0.59) | 2.28 (0.83) | 7.078 | .010 | LI > NL-Match | .005 |
| | | | | | | LI > NL-Compare | .024 |
| Verbal mediation | 2.61 (0.85) | 2.11 (0.58) | 2.33 (0.77) | 2.053 | .208 | | *ns* |
| Flexibility | 3.17 (1.04) | 2.28 (0.75) | 2.17 (0.99) | 6.182 | .015 | LI > NL-Match | .016 |
| | | | | | | LI > NL-Compare | .009 |
| Responsiveness to feedback | 1.89 (0.47) | 1.67 (0.59) | 1.56 (0.51) | 1.859 | .227 | | *ns* |
| Attention | 2.39 (1.15) | 1.67 (0.77) | 1.50 (0.71) | 4.019 | .025 | LI > NL-Match | .030 |
| | | | | | | LI > NL-Compare | .012 |
| Compliance | 2.00 (0.77) | 1.28 (0.58) | 1.28 (0.58) | 7.521 | .010 | LI > NL-Match | .007 |
| | | | | | | LI > NL-Compare | .007 |

*Note*. Numbers in parentheses are standard deviations.

[a]Benjamini and Hochberg (1985) adjusted $p$ values.

Table 4. The full selection model had an $R^2$ of .431, $F(27, 26) = 3.107$, $p = .003$. The forward and stepwise selection models were identical, $R^2 = .500$, $F(4, 49) = 12.270$, $p < .001$. The backward selection model had an $R^2$ of .590, $F(7, 46) = 9.450$, $p < .001$.

The backward multiple-regression model was adopted because it resulted in the most parsimonious model while maximizing the variance. Specifically, the backward model added three variables over the stepwise and forward models (which were identical), significantly increasing the $R^2$. The variables in the model included three of 12 modifiability items (compliance, metacognition, and task orientation), three of 10 posttest story scores (setting, knowledge of dialogue, and complexity of vocabulary), and one of the five SALT-derived story measures (ungrammaticality). These seven scores were entered into the discriminant analyses.

*Discriminant analysis.* Two discriminant analyses were conducted. In the first analysis, we entered LI and NL-Match cases in the first phase and cross-validated the results with the NL-Compare cases. We then repeated the discriminant analysis entering the LI and NL-Compare groups first and cross-validated the results with the original LI and NL-Match groups. Deriving the cut points from two sets of groups and cross-validating those cuts allowed us to examine the robustness of the discriminant function.

The results indicated that the combination of seven scores extracted from the multiple regression demonstrated good classification in the first analysis. First, we tested the equality of the group covariance matrices using Box's *M* statistic. The assumption of equality was met ($p = .081$), and the log determinants were similar (LI = −9.114, NL-Match = −8.312). The overall $\chi^2$ test was significant (Wilks's $\lambda = .368$), $\chi^2(7) = 30.459$, canonical correlation = .795, $p < .001$. This combination of predictors classified 88.9% of the cases accurately with 88.9% sensitivity and 88.9% specificity. The cross-validated classification with the NL-Compare children demonstrated accurate classification of 80.6%, with 88.9% sensitivity and 72.2% specificity.

The final research question concerned replication of the results with a second NL group that was not matched as closely to the LI group. Using the same seven predictors as in the first solution, we reran the discriminant analysis, first entering the LI group with the NL-Compare group, then cross-validating with the LI and NL-Match groups. Examination of Box's *M* indicated that the assumption of equality of covariance matrices was violated ($p = .026$), and the log determinants were dissimilar (LI = −9.114, NL-Match = −5.641). The analysis was rerun with separate covariance matrices in the classification. The new results did not improve the accuracy rate of the discriminant model; thus, the original results are reported. The overall $\chi^2$ test was significant (Wilks's $\lambda = .397$), $\chi^2(7) = 28.170$, canonical correlation = .776, $p < .001$. This combination of predictors classified 94.4% of the cases accurately with 100%

**Table 4.** Correlations and summary statistics from regression models tested.

| Variable | Correlation with language ability | Beta weights from the four models tested | | | |
|---|---|---|---|---|---|
| | | Full | Forward | Backward | Stepwise |
| Anxiety | –.210 | –0.097 | | | |
| Motivation | –.398** | –0.464* | | | |
| Nonverbal persistence | –.340** | 0.266 | | | |
| Task orientation | –.408** | 0.800 | –0.410* | 0.573** | –0.41* |
| Metacognition | –.525** | 1.045** | 0.616** | –0.75** | –0.616** |
| Nonverbal self-reward | –.293* | 0.177 | | | |
| Problem solving | –.447** | –0.051 | | | |
| Verbal mediation | –.245* | 0.141 | | | |
| Flexibility | –.439** | 0.228 | | | |
| Responsiveness to feedback | –.246* | –0.100 | | | |
| Attention | –.399** | –0.059 | | | |
| Compliance | –.477** | 0.468** | 0.333** | –0.378** | –0.333** |
| Setting | .253* | 0.344 | | 0.261* | |
| Character information | .223 | –0.293 | | | |
| Temporal markers | .328** | –0.022 | | | |
| Causal relationships | .156 | –0.138 | | | |
| Complexity of ideas | .129 | –0.290 | | | |
| Complexity of vocabulary | .240* | 0.498* | | 0.214* | |
| Grammatical complexity | .379** | 0.217 | | | |
| Dialogue | .092 | –0.477* | | –0.226 | |
| Creativity | –.083 | –0.022 | | | |
| Episode structure | .269* | 0.027 | | | |
| MLUw | .288* | –0.122 | | | |
| NDW | .236* | 0.247 | | | |
| TNW | .150 | –0.056 | | | |
| Ungrammatical utterances | –.467** | –0.250 | –0.396** | –0.328** | –0.396** |
| Main verbs per utterance | .284* | 0.227 | | | |

*$p \leq .05$. **$p \leq .01$.

sensitivity and 88.9% specificity. The cross-validated classification with the NL-Match children demonstrated accurate classification of 97.2%, with 100% sensitivity and 94.4% specificity.

Under both solutions, the NL-Compare group had more misclassified cases. Recall that this group was matched for language experience and age, but not on IQ and sex. The second solution yielded the best classification for both sets of matched children. The classification function yielded in the discriminant analysis can be used to directly compute classification scores for new observations as follows:

$$\begin{aligned} D = & -1.004 \times \text{Task Orientation} + 1.457 \\ & \times \text{Metacognition} + .809 \times \text{Compliance} - .931 \\ & \times \text{Complexity of Vocabulary Posttest} + .333 \\ & \times \text{Knowledge of Dialog} - .401 \times \text{Setting} + 1.85 \\ & \times \text{Ungrammaticality} - 2.748 \end{aligned}$$

The mean discriminant scores were −1.197 for the NL group and 1.197 for group with LI. Lower scores on the discriminant function were associated with typical development, and high scores were associated with LI (see Table 5).

## Discussion

Accurate differentiation of language difference and LI continues to be a pressing need in assessing bilingual children (Bedore & Peña, 2008; Dollaghan & Horner, 2011; Kohnert, 2010; Thordardottir et al., 2006). Despite substantial progress in identifying diagnostic markers of impairment in Spanish–English bilingual individuals (Armon-Lotem & Walters, 2011; Dollaghan & Horner, 2011; Pearson, 2010), there is a severe shortage of bilingual speech-language pathologists with the knowledge and skills to implement these emerging measures and procedures (American Speech-Language-Hearing Association, 2008). Many speech-language pathologists continue to use English-only standardized tests (Caesar & Kohler, 2007; Williams & McLeod, 2012) and do not feel prepared to make diagnostic decisions about bilingual children. Although best practice procedures include testing in both the first and second language for bilingual persons, there is also a need to systematically document practices in the assessment of English that are potentially informative (Gillam et al., 2013; Paradis et al., 2013).

Children with LIs have difficulties in language learning due in part to inefficiencies in attention and memory (Ellis Weismer et al., 1999, 2000; Gillam et al., 2009; Kohnert et al., 2006). Our approach to DA includes clinician observation of children's strategy use as they are actively engaged in language learning. Such a focus can enable clinicians to systematically observe attention and memory processes as children learn to tell more complete and complex stories. In the present study, we extended this notion to observation of learning in bilingual children's less familiar language. We explored differences in narrative performance of two groups of NL children and children with LI on posttest measures and pretest–posttest patterns. We also compared cognitive and affective strategies used by children with and without LI. Finally, we determined whether DA administered in English could accurately classify bilingual English language learners with and without LI.

### *Storytelling Ability*

There were modest differences between ability groups for DAI total story scores, with children in all three groups scoring higher on the posttest. It appears that two short mediation sessions can result in changes in discourse, even when bilingual English language learners are taught only in English. In contrast, there were no pretest–posttest changes in the productivity and complexity measures derived from the spontaneous stories told in response to wordless picture books, suggesting that the incidental focus on sentence length and complexity had a minimal effect on language productivity. There were however, significant differences between children with and without LI on the DAI story scores and on our grammaticality measure.

The higher posttest scores for the narrative measure are consistent with previous findings of language learning

**Table 5.** A formula table for deriving discriminant scores for the purpose of differentiating between bilingual children with and without hearing impairments.

| Score | Function coefficients | Case 1 scores | Function Coefficient × Score | Case 2 scores | Function Coefficient × Score |
|---|---|---|---|---|---|
| MLO task orientation | –1.004 | 3 | –3.012 | 1 | –1.004 |
| MLO metacognition | 1.457 | 4 | 5.827 | 1 | 1.457 |
| MLO compliance | 0.809 | 3 | 2.428 | 2 | 1.618 |
| DAI complexity of vocabulary | –0.931 | 1 | –0.931 | 1 | –0.931 |
| DAI knowledge of dialogue | 0.333 | 1 | 0.333 | 1 | 0.333 |
| DAI setting | –0.401 | 1 | –0.401 | 2 | –0.802 |
| Proportion of ungrammatical utterances | 1.850 | 0.85 | 1.573 | 0.22 | 0.407 |
| Constant | –2.748 | | –2.748 | | –2.748 |
| *D* = (sum of function coefficients and constant) | | | 3.068 | | –1.671 |

*Note*. MLO = mediated learning observation; DAI = dynamic assessment and intervention.

after mediation (Kramer et al., 2009; Peña et al., 2006). However, our failure to find that children in the two NL groups made greater changes than the children in the LI group is in contrast to previous findings with monolingual first and second graders (Peña et al., 2006) and third graders (Kramer et al., 2009). There are two interrelated possibilities for this pattern of results. Although most preschool- and kindergarten-age children may have relatively good comprehension of stories, not all children consistently produce goal-directed stories that result in a clear consequence. In fact, Stein and Albro (1997) found that only 52% of kindergarten-age children generated stories with goal-directed sequences. It may be that the task of generating a story from a wordless picture book was relatively difficult for the bilingual kindergarten children in our study, whether they had LIs or not. Although both NL and LI groups made some gains, the general difficulty of the task may have restricted the NL groups from making relatively larger gains than the children in the LI group.

Another possibility for the lack of group differences on the posttest language measures relates to the demands of the storytelling task. The narrative language demands of the task may have exceeded the children's English skills. Level of first- and second-language exposure affects vocabulary knowledge (Thordardottir & Brandeker, 2013), processing skills (Roberts, 2012), and storytelling skills (Gutiérrez-Clellen, 2002). Even though children were able to tell simple stories at pretest and posttest, and there were modest differences by ability, difficulty understanding and producing complex English may have hindered the NL children from making larger gains than the children with LI. It is possible that we would have observed pretest–posttest differences if we had used simple, supported story sequences.

### *MLOs*

Children with and without LI did demonstrate differential task and attention patterns, as measured by our MLO. This is consistent with Duinmeijer, de Jong, and Scheper (2012), who found that narrative generation was associated with sustained attention, whereas story recall was associated with working memory. Children with LI scored lower than typically developing children on both of their attention measures. Although the story generation task was challenging for children with and without LIs in our study, clinicians were nonetheless able to observe group differences on a number of behaviors related to modifiability. It is possible that more simple language learning tasks, such as picture description, or shorter story sequences than those used in the current study would be difficult enough to show pretest-to-posttest changes as well as allowing observation of modifiability during learning.

The modifiability observations were highly robust for differentiating between the learning behaviors of the children in the LI and NL groups. These findings are consistent with previous work in DA using a test–teach–retest approach (Kapantzoglou et al., 2012; Kramer et al., 2009; Peña et al., 2006; Ukrainetz et al., 2000). In Peña et al.'s (2007) study of first- and second-grade English-speaking children, a composite score of flexibility and metacognition was sufficient for accurate (93%) classification of children with and without impairment. In the current study, ability group differences were greatest for metacognition, compliance, problem solving, and flexibility (in that order), which was generally consistent with previous findings. Recall, however, that none of these measures, on their own, yielded high classification accuracy. It was only when we combined the modifiability scores with the posttest measures that we were able to achieve high classification accuracy for this group of bilingual kindergarten children. A combination of clinical judgments about children's performance during learning and measures of learning outcomes leads to good diagnostic decisions. DA is a valuable diagnostic technique because it provides a structured way for clinicians to analyze learning processes and learning outcomes.

### *Clinical Implications*

DA has consistently demonstrated good classification accuracy when applied to differentiation of LI in children from diverse cultural and linguistic backgrounds, yet clinicians cite a lack of time and lack of training as factors for not using DA in practice (Burns, 1996; Deutsch & Reynolds, 2000; Hammer, Detwiler, Detwiler, Blood, & Dean Qualls, 2004; Hasson & Joffe, 2007; Moore-Brown, Hueta, Uranga-Hernandez, & Peña, 2006). DA using the test–teach–retest approach can take between one (e.g., Kapantzoglou et al., 2012) and five separate sessions (Stubbe Kester, Peña, & Gillam, 2001). The DA presented here was conducted over three sessions with a total time of approximately 1 hr and 10 min. Transcription of the narrative samples and coding took an additional hour of time.

Consistent with Peña et al. (2006), the posttest scores were used in the analysis, providing children with the best opportunity to benefit from practice. Although the pretest narrative may serve as a good starting point for the assessment, it may not be as informative for clinical decision making, so it would not be necessary to transcribe and score it, which cuts transcription and coding time in half. In addition, only those areas contributing to the discriminant function need to be scored (e.g., proportion of ungrammatical utterances from the narrative samples and setting, dialogue, and complexity of vocabulary scores from the DAI). In addition, only three of the original 12 items from the MLO (compliance, metacognition, and task orientation) entered into the discriminant function. Reducing the number of measures and observations entered into analysis would further reduce the time it takes to complete a DA.

The formula provided earlier can be used to derive a discriminant score, *D*. We provide two examples in Table 5. Each of the seven scores is multiplied by its corresponding function coefficient (see Function Coefficient × Score columns). The resulting dividends and the constant are summed to calculate *D* (see bottom row of table). *D* scores are compared to the mean discriminant scores of −1.197 for the typically developing group and 1.197 for the LI

group. Child 1 has a *D* score of 3.068. This score is indicative of LI. Child 2 has a *D* score of −1.671, indicating typical development.

A possible limitation in generalizing the results from this study to actual clinical practice is that all the clinicians were bilingual, with several years of experience working with bilingual children. Although they did not use Spanish during the DA, their judgments of language ability were likely to be related to their knowledge of other bilingual children with and without LI. Monolingual speech-language pathologists may vary much more in their experience and comfort in making the kinds of clinical decisions about modifiability during teaching sessions (Hammer et al., 2004; Roseberry-McKibbin, Brice, & O'Hanlon, 2005).

## Summary

In this study, we explored the classification accuracy of DA with a group of Spanish–English bilingual children, many of whom were in the early stages of English language learning. Specifically, the children in this study represented a range of English exposure from 20% to 80%, with first age at English exposure ranging from birth to age 4. The finding of high classification accuracy for this group of children indicates that it is possible to make a reasonably accurate judgment of language ability from observations of bilingual children with a range of experiences as they are engaged in a language learning task. Although we focused on Spanish–English bilingual children in this study, it is likely that this procedure could be applied to children who speak other languages. However, note that we did not test children with less than 1 year of cumulative exposure to English and a minimum of 20% current exposure to English. Children with less exposure to English might not show the same degree of responsiveness to instruction in their second language as evidenced by the children in this study.

It is important to continue to explore the role of English language identification measures for children who are still in the process of learning English. Although we support dual-language assessment for the purpose of diagnosis of LI for bilingual children, it is not always feasible. There are few psychometrically sound measures of language development in languages other than English and few bilingual clinicians. Many of the available measures for other languages are normed on monolingual respondents (McLeod & Verdon, 2014). However, bilingual children may not demonstrate the same configuration of skills as do their monolingual peers, on whom these measures have been normed. On the theoretical side, the accuracy with which examiners are able to rate behaviors related to language processing is intriguing. Clearly, more research is needed to compare direct measures of processing with the behavioral observations that were used in this study. Nonetheless, our results are consistent with the idea that children with LIs have processing difficulties that can be observed while they are engaged in language intervention sessions. Gaining greater insights into the learning behaviors of children with LI has the potential for providing important information about the nature of children's language learning difficulties and their intervention needs. DA appears to be a useful research context for such endeavors.

## Acknowledgments

This research was supported by Grant R01DC007439 from the National Institute on Deafness and Other Communication Disorders to Elizabeth D. Peña. Elizabeth D. Peña and Ronald B. Gillam have a financial interest in the Dynamic Assessment of Narratives, which was administered to participants in this study. We thank all of the interviewers and testers for their assistance with collecting the data and the school districts for allowing us access to the participants.

## References

**American Speech-Language-Hearing Association.** (2008). *2008 schools survey summary report*. Rockville, MD: Author.

**Armon-Lotem, S., & Walters, J.** (2011). An approach to differentiating bilingualism and language impairment. In J. Guendouzi, F. Loncke, & M. J. Williams (Eds.), *The handbook of psycholinguistic and cognitive processes: Perspectives in communication disorders* (pp. 463–488). New York, NY: Psychology Press.

**Baker, C.** (2001). *Fundamentals of bilingual education and bilingualism*. Buffalo, NY: Mutilingual Matters.

**Bedore, L. M., & Peña, E. D.** (2008). Assessment of bilingual children for identification of language impairment: Current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism, 11,* 1–29.

**Benjamini, Y., & Hochberg, Y.** (1985). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B, 57,* 289–300.

**Bohman, T. M., Bedore, L. M., Peña, E. D., Mendez-Perez, A., & Gillam, R. B.** (2010). What you hear and what you say: Language performance in Spanish–English bilinguals. *International Journal of Bilingual Education and Bilingualism, 13,* 325–344. doi:10.1080/13670050903342019

**Bracken, B. A., & McCallum, R. S.** (1998). *Universal Nonverbal Intelligence Test*. Rolling Meadows, IL: Riverside.

**Burns, M. S.** (1996). Dynamic assessment: Easier said than done. In M. G. Luther, E. Cole, & P. J. Gamlin (Eds.), *Dynamic assessment for instruction: From theory to application* (pp. 182–188). North York, UK: Captus Press.

**Burton, V. J., & Watkins, R. V.** (2007). Measuring word learning: Dynamic versus static assessment of kindergarten vocabulary. *Journal of Communication Disorders, 40,* 335–356.

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

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

**Carlson, J.** (1983). *Applications of dynamic assessment to cognitive and perceptual functioning of three ethnic groups*. Riverside, CA: U.S. Department of Education, National Institute of Education Grant, Educational Resources Information Center.

**Carlson, J., & Wiedl, K.** (1980). *Dynamic assessment: An approach toward reducing test bias*. Honolulu, HI: Western Psychological Association.

**Cleave, P. L., Girolametto, L. E., Chen, X., & Johnson, C. J.** (2010). Narrative abilities in monolingual and dual language learning children with specific language impairment. *Journal of Communication Disorders, 43,* 511–522. doi:10.1016/j.jcomdis.2010.05.005

**Deutsch, R., & Reynolds, Y.** (2000). The use of dynamic assessment by educational psychologists in the UK. *Educational Psychology in Practice, 16,* 311–331.

**Dollaghan, C. A., & Horner, E. A.** (2011). Bilingual language assessment: A meta-analysis of diagnostic accuracy. *Journal of Speech, Language, and Hearing Research, 54,* 1077–1088. doi:10.1044/1092-4388(2010/10-0093)

**Duinmeijer, I., de Jong, J., & Scheper, A.** (2012). Narrative abilities, memory and attention in children with a specific language impairment. *International Journal of Language & Communication Disorders, 47,* 542–555. doi:10.1111/j.1460-6984.2012.00164.x

**Ellis Weismer, S., Evans, J., & Hesketh, L. J.** (1999). An examination of verbal working memory capacity in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 42,* 1249–1260.

**Ellis Weismer, S., Tomblin, J. B., Zhang, X., Buckwalter, P., Chynoweth, J. G., & Jones, M.** (2000). Nonword repetition performance in school-age children with and without language impairment. *Journal of Speech, Language, and Hearing Research, 43,* 865–878.

**Gillam, R. B., Montgomery, J. W., & Gillam, S. L.** (2009). Attention and memory in child language disorders. In R. G. Schwartz (Ed.), *Handbook of child language disorders* (pp. 201–215). New York, NY: Psychology Press.

**Gillam, R. B., Peña, E. D., Bedore, L. M., Bohman, T. M., & Mendez-Pérez, A.** (2013). Identification of specific language impairment in bilingual children: Part 1. Assessment in English. *Journal of Speech, Language, and Hearing Research, 56,* 1813–1823.

**Gutiérrez-Clellen, V. F.** (2002). Narratives in two languages: Assessing performance of bilingual children. *Linguistics and Education, 13,* 175–197.

**Gutiérrez-Clellen, V. F., & Kreiter, J.** (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24,* 267–288.

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

**Hadley, P. A., Rispoli, M., Fitzgerald, C., & Bahnsen, A.** (2011). Predictors of morphosyntactic growth in typically developing toddlers: Contributions of parent input and child sex. *Journal of Speech, Language, and Hearing Research, 54,* 549–566. doi:10.1044/1092-4388(2010/09-0216)

**Hammer, C. S., Detwiler, J. S., Detwiler, J., Blood, G. W., & Dean Qualls, C.** (2004). Speech-language pathologists' training and confidence in serving Spanish–English bilingual children. *Journal of Communication Disorders, 37,* 91–108.

**Hasson, N., Camilleri, B., Jones, C., Smith, J., & Dodd, B.** (2013). Discriminating disorder from difference using dynamic assessment with bilingual children. *Child Language Teaching and Therapy, 29,* 57–75.

**Hasson, N., & Joffe, V.** (2007). The case for dynamic assessment in speech and language therapy. *Child Language Teaching and Therapy, 23,* 9–25.

**Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R.** (2012). The etiology of variation in language skills changes with development: A longitudinal twin study of language from 2 to 12 years. *Developmental Science, 15,* 233–249. doi:10.1111/j.1467-7687.2011.01119.x

**Hollingshead, A. A.** (1975). *Four-factor index of social status*. New Haven, CT: Yale University Press.

**Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S.** (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools, 43,* 81–96. doi:10.1044/0161-1461(2011/10-0095)

**Kohnert, K. J.** (2010). Bilingual children with primary language impairment: Issues, evidence and implications for clinical actions. *Journal of Communication Disorders, 43,* 456–473. doi:10.1016/j.jcomdis.2010.02.002

**Kohnert, K. J., Windsor, J., & Yim, D.** (2006). Do language-based processing tasks separate children with language impairment from typical bilinguals? *Learning Disabilities Research & Practice, 21,* 19–29.

**Kozulin, A.** (2002). Sociocultural theory and the mediated learning experience. *School Psychology International, 23,* 7–35.

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

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

**Leonard, L.** (2014). *Children with specific language impairment* (2nd ed.). Cambridge, MA: MIT Press.

**Lidz, C. S.** (1991). *Practitioner's guide to dynamic assessment*. New York, NY: Guilford Press.

**Lidz, C. S.** (2002). Mediated learning experience (MLE) as a basis for an alternative approach to assessment. *School Psychology International, 23,* 68–84.

**Mayer, M., & Mayer, M.** (1971). *A boy, a dog, a frog, and a friend*. New York, NY: Penguin Books.

**McLeod, S., & Verdon, S.** (2014). A review of 30 speech assessments in 19 languages other than English. *American Journal of Speech-Language Pathology*. Advance online publication.

**Miller, J., & Iglesias, A.** (2008). Systematic Analysis of Language Transcripts (English and Spanish Version 9) [Computer software]. Middleton, WI: SALT Software.

**Miller, L.** (2000a). *Bird and his ring*. Austin, TX: Neon Rose Productions.

**Miller, L.** (2000b). *Two friends*. Austin, TX: Neon Rose Productions.

**Miller, L., Gillam, R. B., & Peña, E. D.** (2001). *Dynamic assessment and intervention: Improving children's narrative skills*. Austin, TX: Pro-Ed.

**Minami, M.** (2008). Telling good stories in different languages: Bilingual children's styles of story construction and their linguistic and educational implications. *Narrative Inquiry, 18,* 83–110.

**Minami, M., & McCabe, A.** (1995). Rice balls and bear hunts: Japanese and North American family narrative patterns. *Journal of Child Language, 22,* 423–445.

**Moore-Brown, B. J., Hueta, M., Uranga-Hernandez, Y., & Peña, E. D.** (2006). Using dynamic assessment to evaluate children with suspected learning disabilities. *Intervention in School and Clinic, 41,* 209–217.

**Nicolopoulou, A.** (2002). Peer-group culture and narrative development. In S. Blum-Kulka & C. E. Snow (Eds.), *Talking to adults: The contribution of multiparty discourse to language acquisition* (pp. 117–152). Mahwah, NJ: Erlbaum.

**Paradis, J., Schneider, P., & Duncan, T. S.** (2013). Discriminating children with language impairment among English-language learners from diverse first-language backgrounds. *Journal of Speech, Language, and Hearing Research, 56,* 971–981.

**Pearson, B. Z.** (2010). We can no longer afford a monolingual norm. *Applied Psycholinguistics, 31,* 339–343. doi:10.1017/s014271640999052x

**Peña, E. D., Bedore, L. M., Gutiérrez-Clellen, V. F., Iglesias, A., & Goldstein, B.** (2008). *Bilingual English Spanish Oral Screener: Experimental Version*. Unpublished instrument.

**Peña, E. D., Gillam, R. B., Malek, M., Felter, R., Resendiz, M., Fiestas, C., & Sabel, T.** (2006). Dynamic assessment of children from culturally diverse backgrounds: Applications to narrative assessment. *Journal of Speech, Language, and Hearing Research, 49,* 1037–1057.

**Peña, E. D., Iglesias, A., & Lidz, C. S.** (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10,* 138–154.

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

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

**Records, N. L., & Tomblin, J. B.** (1994). Clinical decision making: Describing the decision rules of practicing speech-language pathologists. *Journal of Speech and Hearing Research, 37,* 144–156.

**Restrepo, M. A.** (1998). Identifiers of predominantly Spanish-speaking children with language impairment. *Journal of Speech, Language, and Hearing Research, 41,* 1398–1411.

**Roberts, L.** (2012). Individual differences in second language sentence processing. *Language Learning, 62*(Suppl. 2), 172–188. doi:10.1111/j.1467-9922.2012.00711.x

**Roseberry-McKibbin, C., Brice, A., & O'Hanlon, L.** (2005). Serving English language learners in public school settings: A national survey. *Language, Speech, and Hearing Services in Schools, 36,* 48–61.

**Squires, K. E., Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., Bohman, T. M., & Gillam, R. B.** (2014). Story retelling by bilingual children with language impairments and typically developing controls. *International Journal of Language & Communication Disorders, 49,* 60–74.

**Stein, N. L., & Albro, E. R.** (1997). Building complexity and coherence: Children's use of goal-structured knowledge in telling stories. In M. G. W. Bamberg (Ed.), *Narrative development: Six approaches* (pp. 5–44). Mahwah, NJ: Erlbaum.

**Stubbe Kester, E., Peña, E. D., & Gillam, R. B.** (2001). Outcomes of dynamic assessment with culturally and linguistically diverse students: A comparison of three teaching methods. *Journal of Cognitive Education and Psychology, 2,* 42–59.

**Thordardottir, E.** (2010). Towards evidence-based practice in language intervention for bilingual children. *Journal of Communication Disorders, 43,* 523–537. doi:10.1016/j.jcomdis.2010.06.001

**Thordardottir, E., & Brandeker, M.** (2013). The effect of bilingual exposure versus language impairment on nonword repetition and sentence imitation scores. *Journal of Communication Disorders, 436,* 1–16. doi:10.1016/j.jcomdis.2012.08.002

**Thordardottir, E., Rothenberg, A., Rivard, M.-E., & Naves, R.** (2006). Bilingual assessment: Can overall proficiency be estimated from separate measurement of two languages? *Journal of Multilingual Communication Disorders, 4,* 1–21. doi:10.1080/14769670500215647

**Tomblin, J. B.** (2006). A normativist account of language-based learning disability. *Learning Disabilities Research & Practice, 21,* 8–18.

**Tomblin, J. B., Records, N. L., & Zhang, X.** (1996). A system for the diagnosis of specific language impairment in kindergarten children. *Journal of Speech and Hearing Research, 39,* 1284–1294.

**Tzuriel, D.** (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12,* 385–435.

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

**Wang, Q., & Leichtman, M. D.** (2000). Same beginnings, different stories: A comparison of American and Chinese children's narratives. *Child Development, 71,* 1329–1346.

**Werner, E. O. H., & Kresheck, J.** (1983). *SPELT-II: Structured Photographic Expressive Language Test*. DeKalb, IL: Janelle Publications.

**Wiig, E. H., Secord, W., & Semel, E.** (2005). *Clinical Evaluation of Language Fundamentals–Preschool* (*Second Edition*). San Antonio, TX: The Psychological Corporation.

**Williams, C. J., & McLeod, S.** (2012). Speech-language pathologists' assessment and intervention practices with multilingual children. *International Journal of Speech-Language Pathology, 14,* 292–305. doi:10.3109/17549507.2011.636071