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

Elizabet 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, & Bahnson, 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...
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 modifiability, Peña, Quinn, and Iglesias (1992), narrative production of language tasks such as word learning (Peña, Iglesias, & Lidz, 2001; Peña, Quinn, & Iglesias, 1992), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006). 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, Iglesias, & Lidz, 2001; Peña, Quinn, & Iglesias, 1992), narrative production (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006), and categorization (Kramar, Mallett, Schneider, & Hayward, 2009; Peña et al., 2006). 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 task 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 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 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 (Brucken & 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 bilingual 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
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, 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

### Table 1. Demographic data.

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

**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.
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 responsibility 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, transdisclosure, 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, char-
acter information, and temporal order of events), ideas and
language (grammatical complexity, complexity of idea, cre-
vativity) and episode structure. A second coder independently
scored 32 pairs of pretest and posttest stories (64 stories to-
total) that were randomly selected from the 167 pairs of stories
that had been coded. Point-to-point agreement for the indi-
vidual 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 reliabil-
ity 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 mul-
tivariate Wilks’s λ results, which do not require sphericity
because they are based on difference scores that are com-
bined 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, η² = .124; and time, F(1, 51) =
4.722, p = .034, η² = .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
courses, 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 examine MLUw, NDW, TNW, percentage ungrammat-
icity, 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 mini-
mize the false discovery rate, a Benjamini and Hochberg
(1985) adjustment was used to estimate an adjusted signifi-
cant 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 small-
est 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 re-
port 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 sig-
ificant main effect of group, F(2, 51) = 6.544, p = .045,
η² = .124; and time, F(1, 51) = 19.70, 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 imme-
diately 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 compo-
ents 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.

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

*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 \(\times\) 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.

<table>
<thead>
<tr>
<th>Item</th>
<th>Group</th>
<th>Univariate ANOVA</th>
<th>Post hoc</th>
<th>(p^a)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>LI</td>
<td>NL-Match</td>
<td>NL-Compare</td>
<td>(F)</td>
</tr>
<tr>
<td>Anxiety</td>
<td></td>
<td>1.61 (0.50)</td>
<td>1.44 (0.51)</td>
<td>1.33 (0.49)</td>
</tr>
<tr>
<td>Motivation</td>
<td></td>
<td>2.44 (0.71)</td>
<td>1.78 (0.65)</td>
<td>1.83 (0.79)</td>
</tr>
<tr>
<td>Nonverbal persistence</td>
<td></td>
<td>2.28 (0.75)</td>
<td>1.83 (0.92)</td>
<td>1.56 (0.62)</td>
</tr>
<tr>
<td>Task orientation</td>
<td></td>
<td>3.06 (0.94)</td>
<td>2.11 (0.83)</td>
<td>2.28 (1.02)</td>
</tr>
<tr>
<td>Metacognition</td>
<td></td>
<td>3.50 (0.86)</td>
<td>2.33 (0.84)</td>
<td>2.39 (0.98)</td>
</tr>
<tr>
<td>Nonverbal self-reward</td>
<td></td>
<td>2.39 (1.04)</td>
<td>1.89 (0.58)</td>
<td>1.89 (0.68)</td>
</tr>
<tr>
<td>Problem solving</td>
<td></td>
<td>2.94 (0.87)</td>
<td>2.00 (0.59)</td>
<td>2.28 (0.83)</td>
</tr>
<tr>
<td>Verbal mediation</td>
<td></td>
<td>2.61 (0.85)</td>
<td>2.11 (0.58)</td>
<td>2.33 (0.77)</td>
</tr>
<tr>
<td>Flexibility</td>
<td></td>
<td>3.17 (1.04)</td>
<td>2.28 (0.75)</td>
<td>2.17 (0.99)</td>
</tr>
<tr>
<td>Responsiveness to feedback</td>
<td></td>
<td>1.89 (0.47)</td>
<td>1.67 (0.59)</td>
<td>1.56 (0.51)</td>
</tr>
<tr>
<td>Attention</td>
<td></td>
<td>2.39 (1.15)</td>
<td>1.67 (0.77)</td>
<td>1.50 (0.71)</td>
</tr>
<tr>
<td>Compliance</td>
<td></td>
<td>2.00 (0.77)</td>
<td>1.28 (0.58)</td>
<td>1.28 (0.58)</td>
</tr>
</tbody>
</table>

Note. Numbers in parentheses are standard deviations.

\(^a\)Benjamini and Hochberg (1985) adjusted \(p\) values.
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 χ² test was significant
(Wilks’s λ = .368), χ²(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 classifi-
cation 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 co-
variance 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 χ² test
was significant (Wilks’s λ = .397), χ²(7) = 28.170, canonical
correlation = .776, p < .001. This combination of predic-
tors classified 94.4% of the cases accurately with 100%

Table 4. The full selection model had an R² of .431, F(27, 26) =
3.107, p = .003. The forward and stepwise selection models
were identical, R² = .500, F(4, 49) = 12.270, p < .001. The
backward selection model had an R² 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 R². 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 dia-
logue, 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

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

<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation with language ability</th>
<th>Full</th>
<th>Forward</th>
<th>Backward</th>
<th>Stepwise</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anxiety</td>
<td>−.210</td>
<td>−.097</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>−.398**</td>
<td>−.464*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonverbal persistence</td>
<td>−.340**</td>
<td>0.266</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Task orientation</td>
<td>−.408**</td>
<td>0.880</td>
<td></td>
<td>−.410*</td>
<td>0.573**</td>
</tr>
<tr>
<td>Metacognition</td>
<td>−.525**</td>
<td>1.045**</td>
<td>0.616**</td>
<td>−.75**</td>
<td>−.616**</td>
</tr>
<tr>
<td>Nonverbal self-reward</td>
<td>−.293*</td>
<td>0.177</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Problem solving</td>
<td>−.447**</td>
<td>−.051</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal mediation</td>
<td>−.245*</td>
<td>0.141</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexibility</td>
<td>−.439**</td>
<td>0.228</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Responsiveness to feedback</td>
<td>−.246*</td>
<td>−.100</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attention</td>
<td>−.399**</td>
<td>−.059</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compliance</td>
<td>−.477**</td>
<td>0.468**</td>
<td>0.333**</td>
<td>−.378**</td>
<td>−.333**</td>
</tr>
<tr>
<td>Setting</td>
<td>.253*</td>
<td>0.344</td>
<td>0.261*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Character information</td>
<td>.223</td>
<td>−.293</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temporal markers</td>
<td>.328**</td>
<td>−.022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Causal relationships</td>
<td>.156</td>
<td>−.138</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity of ideas</td>
<td>.129</td>
<td>−.290</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Complexity of vocabulary</td>
<td>.240*</td>
<td>0.498*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grammatical complexity</td>
<td>.379**</td>
<td>0.217</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dialogue</td>
<td>.092</td>
<td>−.477*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Creativity</td>
<td>−.083</td>
<td>−.022</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Episode structure</td>
<td>.269*</td>
<td>0.027</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLU</td>
<td>.288*</td>
<td>−.122</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NDW</td>
<td>.236*</td>
<td>0.247</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TNW</td>
<td>.150</td>
<td>−.056</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ungrammatical utterances</td>
<td>−.467**</td>
<td>−.250</td>
<td>−.396**</td>
<td>−.328**</td>
<td>−.396**</td>
</tr>
<tr>
<td>Main verbs per utterance</td>
<td>.284*</td>
<td>0.022</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*p ≤ .05. **p ≤ .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:

\[ D = -1.004 \times \text{Task Orientation} + 1.457 \]
\[ \times \text{Metacognition} + 0.809 \times \text{Compliance} - 0.931 \]
\[ \times \text{Complexity of Vocabulary Posttest} + 0.333 \]
\[ \times \text{Knowledge of Dialog} - 0.401 \times \text{Setting} + 1.85 \]
\[ \times \text{Ungrammaticality} - 2.748 \]

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.

<table>
<thead>
<tr>
<th>Score</th>
<th>Function Coefficients</th>
<th>Case 1</th>
<th>Function Coefficient × Score</th>
<th>Case 2</th>
<th>Function Coefficient × Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLO task orientation</td>
<td>-1.004</td>
<td>3</td>
<td>-3.012</td>
<td>1</td>
<td>-1.004</td>
</tr>
<tr>
<td>MLO metacognition</td>
<td>1.457</td>
<td>4</td>
<td>5.827</td>
<td>1</td>
<td>1.457</td>
</tr>
<tr>
<td>MLO compliance</td>
<td>0.809</td>
<td>3</td>
<td>2.428</td>
<td>2</td>
<td>1.618</td>
</tr>
<tr>
<td>DAI complexity of vocabulary</td>
<td>-0.931</td>
<td>1</td>
<td>-0.931</td>
<td>1</td>
<td>-0.931</td>
</tr>
<tr>
<td>DAI knowledge of dialogue</td>
<td>0.333</td>
<td>1</td>
<td>0.333</td>
<td>1</td>
<td>0.333</td>
</tr>
<tr>
<td>DAI setting</td>
<td>-0.401</td>
<td>1</td>
<td>-0.401</td>
<td>2</td>
<td>-0.802</td>
</tr>
<tr>
<td>Proportion of ungrammatical utterances</td>
<td>1.850</td>
<td>0.85</td>
<td>1.573</td>
<td>0.22</td>
<td>0.407</td>
</tr>
<tr>
<td>Constant</td>
<td>-2.748</td>
<td>0.85</td>
<td>-2.748</td>
<td>0.22</td>
<td>-2.748</td>
</tr>
<tr>
<td>( D = (\text{sum of function coefficients and constant}) )</td>
<td>3.068</td>
<td>1.671</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 Duimmeijer, de Jong, and Schep (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 $\times$ 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 LI 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


Peña et al.: Dynamic Assessment Accurately Identifies LI 2219


