JSLHR

Research Article

Dynamic Assessment of Narratives: Efficient, Accurate Identification of Language Impairment in Bilingual Students

Douglas B. Petersen,[a] Helen Chanthongthip,[b] Teresa A. Ukrainetz,[a] Trina D. Spencer,[c] and Roger W. Steeve[a]

Purpose: This study investigated the classification accuracy of a concentrated English narrative dynamic assessment (DA) for identifying language impairment (LI).
Method: Forty-two Spanish–English bilingual kindergarten to third-grade children (10 LI and 32 with no LI) were administered two 25-min DA test–teach–test sessions. Pre- and posttest narrative retells were scored in real time. Using a structured intervention approach, examiners taught children missing story grammar elements and subordination. A posttest was administered using a parallel story.
Results: Four classification predictors were analyzed: posttest scores, gain scores, modifiability ratings, and teaching duration. Discriminant function analysis indicated that an overall modifiability rating was the best classifier, with 100% sensitivity and 88% specificity after 1 DA session and 100% sensitivity and specificity after 2 sessions. Any 2 combinations of posttest scores, modifiability ratings, and teaching duration for just 1 session resulted in sensitivity and specificity rates over 90%. Receiver operating characteristic analyses were used to identify clinically usable cutoff points. Post hoc exploration indicated that similar results could be obtained after only one 5–10-min teaching cycle, potentially further abbreviating the DA process.
Conclusion: Concentrated English narrative DA results in high classification accuracy for bilingual children with and without LI. This efficient version of DA is amenable to clinical use.

Children with language impairment (LI) must be identified to receive individualized educational support, including language intervention. Despite considerable evidence concerning the lack of sensitivity and specificity in the identification of LI for children from culturally and linguistically different backgrounds, and federal regulations allowing alternative measures, norm-referenced testing continues to be almost the solely acceptable method of identifying children for services (Caesar & Kohler, 2007; Figueroa & Newsome, 2006; Gandara, 2010; National Research Council, 2002). This study investigates the validity of one promising assessment measure, *dynamic assessment* (DA), with modifications that may make it more acceptable for clinical use.

By examining the mediated learning or teaching process, rather than the independent product of prior learning experiences, DA avoids many of the sources of bias and other classification issues associated with conventional norm-referenced testing. However, despite strong evidence of identification accuracy across a variety of language targets, protocols, and cultural-linguistic groups, and repeated recommendations for wider use, DA has not been adopted in clinical practice. Reasons for lack of adoption of this seemingly highly desirable assessment procedure may include the lack of standardized protocols and materials, the length of training, administration and scoring time, the absence of validated cut points to indicate acceptable versus impaired performance, or the subjective nature of the modifiability ratings often used to gauge student learning (Hasson & Joffe, 2007).

The current study addresses several of these concerns. This study investigates the validity of a test–teach–test DA of narrative language administered in English, concentrated into a single session with a real-time scoring procedure. This DA significantly reduces the time requirement compared to conventional DA, with the intention of making it more amenable to clinical use. The procedures

[a]University of Wyoming, Laramie
[b]Synertx Rehabilitation, Great Falls, Montana
[c]Northern Arizona University, Flagstaff
Correspondence to Douglas B. Petersen: dpeter39@uwyo.edu

Editor: Sean Redmond
Associate Editor: Nicole Terry
Received December 10, 2015
Revision received May 24, 2016
Accepted October 24, 2016
https://doi.org/10.1044/2016_JSLHR-L-15-0426

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

and materials have been used and refined in other controlled studies and clinical applications (Petersen et al., 2014; Petersen, Gillam, Spencer, & Gillam, 2010; Spencer & Slocum, 2010) and use materials readily available to clinicians (Petersen & Spencer, 2010, 2012, 2014; Spencer & Petersen, 2012).

DA for Difference Versus Disorder

DA focuses on current learning rather than the product of prior learning. This judgment of a child's ability to learn is often referred to as *modifiability*. Centered on Vygotsky's (1978) theory of cognitive development, DA seeks to determine the child's "zone of proximal development" (Gutiérrez-Clellen & Peña, 2001; Lidz & Peña, 1996) —that is, DA is used to determine the magnitude of the difference between the level of performance a child can reach unassisted and the level that is reached with mediation or teaching support. A large zone indicates high modifiability and strong potential to learn.

The most common format of DA is *test–teach–test*. During the first phase, a child is administered a brief test on the relevant content to obtain an initial measure of his or her independent performance. In the teaching phase, the examiner provides a brief period of instruction. The teaching should address both the target language skills and the associated learning behaviors, such as transfer of learning, response to prompts, sustaining attention, reflective responding, and dealing with challenge. Following the teaching phase, the child is retested using the same or an alternate parallel form of the initial test. Test scores are examined in two ways: gains in performance from pre- to posttest and level of posttest score. The expectation of DA is that the larger the gain and the higher the final test score, the greater the learning ability. In addition to pretest, posttest, and gain scores, the DA yields an index of response to instruction, or modifiability, on the basis of ratings of how well the child attends, responds, integrates, and applies the skills taught. Modifiability items are designed to tap distinctive contributions to learning potential, including responsiveness to prompts, degree of transfer, attention, ease of acquisition, frustration, disruptions, and the ease with which the examiner is able to obtain optimal response from the child (Gutiérrez-Clellen & Peña, 2001; Lidz & Pena, 1996).

Judgments of learning potential are made by examining how much change occurs between pre- and posttest scores, what learning behaviors were exhibited, and how much effort the examiner has to expend in teaching the child. A small change in performance, a poor posttest score, poor learning behaviors, and a lot of examiner effort is considered to indicate a language disorder, and the inverse indicates typical development. By examining the process of learning rather than the product of prior learning experiences, DA circumvents many differences in life experiences, language, and culture. Because DA is considered to tap inherent learning ability, not current language knowledge, it can be administered in the child's second language.

Peña and Iglesias (1992) conducted a seminal study of DA for differentiating difference versus disorder in the field of speech-language pathology. Fifty Puerto Rican and African American 3–4-year-old children attending Head Start were identified as having typical or low language development on the basis of classroom observation and reports from teachers and parents. Both groups of children had scored poorly on a traditional single-word expressive labeling test. Many of the errors on the expressive vocabulary test involved giving descriptions rather than labels, so the investigators targeted the principle of labeling for DA mediation. The children received two individual mediated learning sessions on using "special names" in a test–teach–test format. Following mediation, ratings of learner modifiability and standardized test scores differentiated the higher and lower language ability groups. The DA correctly classified 92% of the participants, whereas the static, traditional vocabulary assessment correctly classified only 37% of the participants.

Peña and Iglesias (1992) provided early support for DA. Since then, a strong body of evidence has emerged supporting use of DA for determining language differences versus disorders for individuals from culturally and linguistically different backgrounds (e.g., Hasson, Camilleri, Jones, Smith, & Dodd, 2012; Hasson, Dodd, & Botting, 2012; Kramer, Mallett, Schneider, & Hayward, 2009; Lidz & Peña, 1996; Peña, Gillam, & Bedore, 2014; Peña et al., 2006; Peña & Iglesias, 1992; Ukrainetz, Harpell, Walsh, & Coyle, 2000). A recent study by Hasson, Camilleri, et al. (2012) examined a multipronged DA procedure that examined children's ability to learn vocabulary, sentence structure, and phonology. In the study, group performance of 12 bilingual children receiving speech-language services and 14 bilingual children not in therapy were distinguished by amount of prompting or posttest performance across the three areas of communication.

DA Applied to Narrative Language

One language area that has shown considerable potential for DA is *narrative*. Narratives provide a multifaceted skill context: There is almost always something on which children can improve, whether it is vocabulary, linguistic grammar, cohesion, story grammar, or story art. Furthermore, narrative language has been shown to be a stronger predictor of later language and literacy difficulties than word- and sentence-level tasks (Bishop & Edmundson, 1987; Fazio, Naremore, & Connell, 1996; Wetherell, Botting, & Conti-Ramsden, 2007).

Another attractive feature of DA of narratives is that it can accomplish dual goals of identifying LI and providing specific direction for intervention. The few skills that are addressed during the teaching phase of DA are unlikely to be mastered in that brief time, so they can become later intervention goals. Other identified language needs can be addressed in treatment following DA. Furthermore, these teaching sessions often reveal other learning behaviors that are in need of improvement, such as waiting to respond,

applying a skill to a new task, and evaluating one's own performance. Stock-Shumway (1999) found that in investigating DA as a kindergarten screener, conventional language and articulation could be obtained during the teaching phase in addition to the learner descriptions. Thus, a DA of narratives not only can help identify language disorders, but can also provide language and learning goals for subsequent intervention.

Investigations of DA of narratives have generally indicated high classification accuracy (Kramer et al., 2009; Peña et al., 2006, 2014). Peña et al. (2006) examined the classification accuracy of a DA of narratives using procedures outlined by L. Miller, Gillam, and Peña (2001). The researchers examined the DA classification accuracy on 71 children with and without LI. The participants were first and second graders of European American, African American, and Latino American backgrounds. The children with LI were identified by having at least three of four indicators: teacher concern, parent concern, observation of spoken language errors during peer interaction, or a score more than 1 *SD* below the mean on an omnibus spoken language test. Ratings of modifiability were significantly stronger for the children with typical language. For these children of diverse backgrounds, Peña, et al. (2006) found that the modifiability score was the single best indicator of LI, with an overall classification accuracy of 93% sensitivity and 82% specificity. Using a combination of modifiability scores with posttest scores, Peña et al. (2006) achieved 100% sensitivity and 100% specificity. Post hoc analysis showed a similar pattern of performance among the three ethnic groups.

Kramer et al. (2009) replicated Peña et al. (2006) on a sample of 17 third-grade First Nation (i.e., Native Canadian Indian) children with and without LI. Using discriminant function analysis with modifiability and gain scores, sensitivity was 100% and specificity was 92%. Similar to Peña et al. (2006), this investigation revealed extremely high classification accuracy using DA of narration.

In a recent follow-up study, Peña et al. (2014) used the same English narrative DA procedures with 18 bilingual children with LI; 18 bilingual children with normal language development matched on age, sex, language experience, and IQ; and an additional 18 bilingual children with normal language development matched on only age and language experience. The DA procedure was conducted over three sessions across a 7- to 14-day period, with the first session comprising the pretest and the first 30-min intervention, the second session comprising the second 30-min intervention, and the third session the posttest. Pre- and posttest narrative assessments were audio-recorded, transcribed, and analyzed using the Systematic Analysis of Language Transcripts (SALT; J. F. Miller & Iglesias, 2012) with additional coding for grammaticality and sentential complexity. This DA took place over three sessions with a total time of approximately 70 min, with an additional hour for transcription and coding. Results of Peña et al. (2006) indicated that DA, using a combination of examiner ratings of modifiability, posttest narrative scores, and posttest narrative language ungrammaticality, yielded 81%–97% classification accuracy.

Improving the Efficiency of DA

DA has high diagnostic accuracy and can be used to inform intervention. There is one commercial measure available, that of L. Miller et al. (2001). DA has been explained in the clinical literature (e.g., Elliott, 2003; Gutièrrez-Clellen & Peña, 2001; Hasson & Joffe, 2007; Laing & Kamhi, 2003; Lidz & Peña, 1996) and is typically included in any standard textbook on language assessment, particularly for children with cultural and linguistic differences (e.g., Haynes & Pindzola, 2007). Despite its high visibility, DA is not regularly used in the clinical setting, even for children from culturally and linguistically diverse backgrounds. Hasson and Joffe (2007) discuss reasons for the lack of use of DA by speech-language pathologists (SLPs) in the United Kingdom. No nationwide survey could be located for the United States, but there is no evidence to indicate a different level of adoption in the United States. Caesar and Kohler (2007) conducted a survey of the assessment practices of Michigan SLPs for bilingual children. The survey revealed wide variability in procedures and measures, with a high reliance on English norm-referenced tests. Although a few respondents reported using interpreters, non-English measures, language sampling, and observation, none of the respondents reported using DA.

One way to improve clinical acceptability would be to shorten the DA process. In previous research, the DA assessment and teaching phases have spanned three to four sessions (Peña et al., 2006, 2014). Three to four days of administration in addition to, for narrative DA, the time-consuming task of transcribing and analyzing the narrative samples, affects the efficiency and clinical utility of DA. It is not surprising that with such labor-intensive procedures, SLPs continue to use the more biased norm-referenced tests that can be administered and scored in a single evaluation session.

In efforts to shorten and simplify the process, there has been some investigation of the most predictive components of DA. Modifiability scores have consistently been found to have better classification accuracy than posttest or gain scores (Peña & Iglesias, 1992; Petersen, Allen, & Spencer, 2016; Petersen & Gillam, 2013; Ukrainetz et al., 2000). In terms of number of teaching sessions, Peña, Resendiz, & Gillam, (2007) and Ukrainetz et al. (2000) both found that modifiability scores from two 20-min teaching sessions were more predictive than those from just one session. Stock-Shumway (1999) used a teach-only format, with two 20-min small-group phoneme segmentation teaching sessions in which each child's overall modifiability, language, and articulation skills were rated. This format kept the total administration time no longer than that required for conventional individual screenings. Although beneficial information was obtained, this study had some methodological issues that prevented a determination of the accuracy of this teach-only group format.

Another possibility to reduce the time required for DA is to retain the test–teach–test format, but abbreviate the teaching and testing procedures. Petersen and Spencer (2010, 2012, 2014) have developed brief, reliable narrative intervention and assessment procedures, versions of which have been examined in a variety of controlled investigations (Petersen et al., 2014; Petersen & Spencer, 2010; Petersen, Thompsen, Guiberson, & Spencer, 2015; Spencer, Kajian, Petersen, & Bilyk, 2014; Spencer, Petersen, Slocum, & Allen, 2014; Spencer & Slocum, 2010; Weddle, Spencer, Kajian, & Petersen, 2016). The evidence suggests that DA can be significantly shortened and simplified for clinical use while retaining its power to identify LI in culturally and linguistically diverse children.

Classification Cut Points for DA

In addition to a more efficient DA, SLPs need a standard against which to make clinical decisions. Norm-referenced tests have a normative sample against which a test taker's performance can be compared to determine whether performance is "within the average range," "below average," or "extremely below average." Commonly, *cut points* for test scores, typically expressed as a percentile or standard deviations below the mean, are required to qualify for services in the current education climate (Betz, Eickhoff, & Sullivan, 2013). These cut points can be set to optimize classification of impaired versus typically developing (TD) on the basis of the characteristics of each test. However, it is far more common to use generally accepted cut points with little reference to the match to individual tests or how scores match with performance in communicative activities (Ebert & Scott, 2014; McFadden, 1996; Spaulding, Szulga, & Figueroa, 2012). These generally accepted conventions for judgments of LI and eligibility for SLP services vary from the 16th percentile (−1 *SD*) to the 2nd percentile (−2 *SD*).

Although DA can have purposes that do not rely on standardization or classification accuracy (see Elliott, 2003, for a detailed discussion), there is a patent need for an assessment approach that can be used for classification purposes that is less linguistically and culturally biased than traditional norm-referenced assessments (Lopez, 1997). For DA to be used for impairment and eligibility decisions, classification cut points are required. For DA, the cut points could be achievement of a certain posttest performance, a certain raw score or standard score gain at posttest, or a certain modifiability rating. Most research studies on DA determine cut points for evaluating the classification accuracy of its particular DA task. However, with the wide variation in DA formats and the lack of clinical usage, there are no generally accepted cut points for DA. Although SLPs may value DA in the abstract, and may even use versions of DA informally to guide treatment planning (such as testing for stimulability, taking scaffolding data, or noting learner characteristics), without a standard for what is deficient versus acceptable DA performance, clinicians cannot use DA in impairment and eligibility decisions (Gipps, 1999). An additional benefit of cut points is that it reduces scores and ratings to a dichotomous judgment of plus–minus impairment instead of a multipoint rating or scoring scale. This more simple, binary judgment is likely to be easier to learn and more reliable to use. For DA to be clinically useful, easily interpretable empirically based cut points are needed.

The Current Study

The current study investigates several refinements of DA aimed at improving its efficiency and clinical appeal for identifying LI. The purpose of this study was to determine the classification accuracy of a concentrated DA format for bilingual Spanish-English children using materials and procedures available to clinicians, and conducted in English with real-time scoring and learning target selection.

The DA investigated in the current study involved two 25-min test–teach–test sessions conducted in English, with real-time scoring of the pretest and posttest narrative retells, systematic narrative instruction, and modifiability ratings. Posttest scores, pre- to posttest gains, and two modifiability ratings were compared with respect to classification accuracy. The hypothesis for this study is that, consistent with past research, the modifiability ratings for this novel concentrated, real-time narrative DA would be more accurate than the other potential DA predictors. Reports on interrater reliability for the modifiability scales used in prior narrative DA research has been absent (e.g., Peña et al., 2006, 2014). Because the modifiability rating scale is essentially subjective, face validity of the DA process could be negatively impacted. Therefore, inter- and intrarater reliability was carefully assessed in this study.

Furthermore, this study examined another indicator of responsiveness to teaching, namely how long it takes to teach a child with an inherent language learning impairment versus one who only lacks experience with the teaching target, by timing the first cycle of instruction in the first DA session. The hypothesis for this question was that the first cycle would be sufficient to distinguish types of language learners. Last, in addition to determining optimal combinations of indicators, to expand the clinical utility of DA, cut points that best separated typical versus impaired performance were identified on each predictive indicator using receiver operating curve (ROC) analysis. No hypothesis was set for this exploratory question.

In the current study, for Spanish- and English-speaking kindergartners to third graders previously identified with or without LI, the following research questions were investigated:

1. Does a novel concentrated DA of narratives in English differentiate between bilingual children with and without LI?
2. For this concentrated DA, which of the conventional indicators of posttest scores, gain scores, and modifiability ratings, across one versus two DA sessions, most parsimoniously contribute to classification accuracy?

3. Does a timing of the first teaching cycle duration accurately classify children, and does it provide any additional explanation of variance beyond test scores and modifiability ratings?
4. What are cut points to distinguish typical versus impaired on the most predictive indicators on this DA task?

Method

Participants

Participants were recruited from a large urban school district in the mountain west. Researchers contacted the principals and SLPs of three elementary schools requesting contact with the parents of all children from kindergarten to third grade who were Hispanic with at least minimal proficiency in Spanish and English. Eighty-four children returned signed permission forms. The principals, teachers, and SLPs confirmed that the participating children were bilingual to some degree. Of those 84 children, 17 had an Individualized Education Plan (IEP) for language services and 67 had no identification of LI. The 84 children were randomly assigned to participate in the current study or in another study being conducted by the investigators at the time. As a result, 32 children were randomly selected from the participants without IEPs, and 10 children were randomly selected from the 17 participants with IEPs to participate in the current study. Participants were between the ages of 6;4 and 9;6 (years;months, mean of 7;7).

Parent questionnaires provided information on mother's education, prior preschool attendance, eligibility for free or reduced lunch, and language use and exposure in the home. Complete information was available for all participants except for mother's level of education and preschool attendance. The items completed by a parent of every participant are summarized in Table 1. All the respondents reported that one or more parents spoke Spanish in the home at least 1 hr a day in the child's presence, and that the child could speak at least some English and some Spanish. Language sample analyses of English and Spanish narrative retells were used to confirm language proficiency. If English and Spanish performance on any one of mean length of utterance (MLU), total number of words (TNW), and number of different words (NDW) was within 1 *z* score of each other, then a student was classified as *balanced bilingual*. Twenty-four (57%) of the 42 participants were judged to be balanced bilingual, four children (10%) were bilingual Spanish dominant, and 14 children (33%) were bilingual English dominant. The families were predominantly low socioeconomic status, with 93% qualifying for free or reduced lunch and 72% with maternal education less than high school. Although not statistically significant, the group of children with LI had a smaller percentage of female participants, more third grade participants, fewer second grade participants, and more English-dominant bilingual participants in comparison to the group of children with typical language.

To determine the accuracy of DA classification, a *true typical* versus *true impaired* reference point is needed. To be classified as true impaired, children had to meet four requirements. First, a child had to be currently receiving language services in school, on the basis of an IEP. Second, a bilingual, native Spanish-speaking SLP had to have been involved in the eligibility decision. Third, a child had to perform more than 1 *SD* below the mean in both languages compared to a bilingual story retell database (J. F. Miller & Iglesias, 2012) on at least one of three indicators (MLU, TNW, NDW) when retelling a model story, on the basis of the wordless storybook *Frog, Where Are You?* (Frog Retell; Mayer, 1969). Last, written or verbal confirmation of LI status was required from at least one parent or teacher with no parent or teacher disagreeing with the judgment. A native Spanish-speaking SLP was also involved in the eligibility decision for students classified as true typical. These students could not have an IEP for language services, had to have scores higher than −1 *SD* of the mean on MLU, TNW, and NDW on the Frog Retell, and parents and teachers could not have concerns about the student's language.

Procedures

DA Overview

The entire procedure occurred in 3 days for each participant. Following administration of the Frog Retells, DA was administered over the next 2 days. Trained examiners, blind to the language status of the participants, administered the Frog Retell and DA. All participants received one DA session of approximately 25 min on one day (S1) and another the next day (S2). Each 25-min DA session comprised: (a) a pretest narrative retell, (b) a narrative retell teaching phase, and (c) a posttest narrative retell. The pre- and posttest narrative retells and modifiability ratings were scored during the sessions. For the teaching phase, examiners cycled one to four times through a brief set of structured steps targeting individualized story grammar and adverbial subordinate clauses (see DA Teaching Phase). The story used for the pretest was used during the first teaching cycle. Thereafter, different stories were used in each teaching cycle and for the posttest. The two DA sessions (S1 and S2) used different sets of stories. All Frog Retell and DA sessions were audio-recorded.

Frog Retell Language Sample

Using the wordless storybook *Frog, Where Are You?* (Mayer, 1969), a narrative retell sample (Frog Retell) was collected in English and Spanish on the day prior to the DA administration to determine language dominance and LI. The order of languages used for the Frog Retells was randomly counterbalanced across participants. Examiners read a script from the SALT (J. F. Miller & Iglesias, 2012) manual in either English or Spanish while showing the child corresponding pictures from the wordless picture book. Following the modeled story, the child was given the wordless picture book and asked to retell the story in

Table 1. Participant demographic information.

Characteristics	Total participants (*n* = 42)		LI (*n* = 10)		TD (*n* = 32)	
	n	%	*n*	%	*n*	%
Kindergarten	7	17	2	20	5	16
1st grade	14	33	4	40	10	31
2nd grade	12	29	1	10	11	34
3rd grade	9	21	3	30	6	19
Female	19	45	3	30	16	50
Mother education < high school	23	72	6	60	17	77
Attended preschool	32	100	8	100	24	100
Free or reduced lunch	39	93	9	90	30	94
English dominant bilingual	14	33	4	40	10	31
Spanish dominant bilingual	4	10	0	0	4	13
Balanced bilingual	24	57	6	60	18	56

Note. LI = language impairment; TD = typically developing.

that same language. Children then had a brief break from storytelling for 5–10 min, which often entailed moving to a different examiner at a different testing location, after which the examiner administered the Frog Retell in whichever language not yet sampled.

Each English and Spanish Frog Retell sample was audio-recorded, transcribed, segmented, and analyzed by trained bilingual research assistants using the SALT software (J. F. Miller & Iglesias, 2012). SALT-derived MLU, NDW, and TNW were compared to the SALT Bilingual Spanish/English Story Retell Reference Databases. This database contains Spanish and English retells of *Frog, Where Are You* from over 2,000 bilingual kindergarten through third grade children.

DA Pre- and Posttests.

In the pre- and posttest phases of the DA sessions, the examiner told the child a brief story in English. The stories used for the testing and teaching were structured with elaborated episodes and adverbial subordination similar to those used in the kindergarten-level Narrative Language Measures (NLM; Petersen & Spencer, 2010, 2012). All the stories were parallel in length, story grammar features, and language complexity (see Appendix A for an example story). After a pretest story was read to the participant, the examiner asked the child to retell the story. While the child retold the story, the examiner scored the retell in real time. The total possible score was 33 points: (a) 18 points for presence and quality of each of nine story grammar elements (character, setting, problem, emotion, plan, attempt, consequence, ending, and ending emotion); (b) 10 points for up to one occurrence of *then* and three occurrences of each of *because, when,* and *after;* and (c) 5 points for complexity of episodic structure (e.g., initiating event, attempt, consequence).

DA Teaching Phase

The teaching phase followed the individualized narrative intervention procedures used in previous studies (cf. Petersen et al., 2014; Spencer, Petersen, et al., 2014). Intervention was provided in English. There were four steps to a teaching cycle (see Table 2). Each cycle started with the examiner reading aloud an unfamiliar story similar in structure to the pre- and posttest stories, with clear story grammar elements and adverbial subordinate clauses. The examiner then helped the child retell the story using preset verbal prompts, illustrations, and colored icons representing the main story grammar elements. Examiners targeted any of the story grammar elements and adverbial subordinate clauses that were omitted or poorly represented in the child's narratives.

In terms of teaching targets, all the participants had one adverbial subordinate clause target. A causal subordinate clause (he was scared *because he fell down*) was always targeted unless it was produced without examiner support. In this event, the examiner instead targeted temporal subordinate clauses using the connectives *when* or *after* (e.g., *after he got home* he talked his mom). For story grammar, two or more of nine story grammar elements were always targeted: character (e.g., *John*), setting (activity and location; e.g., *was riding his bike down the street*), problem (e.g., *fell off his bike and got hurt*), emotion (e.g., *was sad*), plan (e.g., *he decided to get help*), attempt (e.g., *he asked*

Table 2. Steps in each dynamic assessment teaching cycle.

Steps	Examiner responsibilities
1. Model narrative	Lay out pictures Model the story Place icons near pictures Name the story grammar parts
2. Retell with pictures and icons	Leave pictures and icons Support child retelling story
3. Retell with icons	Remove pictures Support child retelling story
4. Retell without pictures and icons	Remove icons Support child retelling story

Note. Procedures from *Story Champs* language intervention (Spencer & Petersen, 2012).

his mom for a band aid), consequence (e.g., *the boy received his band aid*), and ending emotion (e.g., *John was happy*).

Visual and verbal support was systematically faded over the cycles to allow for as much independent retelling as possible. Individualized systematic support was provided using the teaching principles outlined in Table 3. A teaching cycle took 5–10 min to administer, with the first cycle being the slowest and most variable. Examiners completed as many cycles as possible in the 15–20-min teaching phase of each session. Examiners never truncated the steps in a teaching cycle, which resulted in sessions occasionally going over the 25 min but no session exceeded 30 min. For the teaching duration indicator, after data collection was completed, audio recordings of the first teaching cycle of the first DA session were timed.

Modifiability

To evaluate modifiability, examiners used a seven-item modifiability rating form immediately after the teaching phases of each DA session (see Appendix B). The items were based on modifiability rating scales used in previous DA research (Peña et al., 2006, 2007; Petersen & Spencer, 2014; Ukrainetz et al., 2000). The items were designed to be easy to score while the session was still underway. Questions 1–6 of the modifiability rating form were focused on how frequently or clearly specific child behaviors occurred during the teaching phase: (a) being responsive to prompts, (b) displaying transfer as the cycles continued, (c) attending to the testing/teaching, (d) ease of teaching, (e) not displaying frustration, and (f) not disrupting the testing session. Each item was rated on a 3-point scale (0–2), with clear examples of 0, 1, and 2 point behaviors offered on the modifiability rating form. Question 7 asked about the child's potential to learn narrative language, or overall modifiability, again on a 3-point scale, with 0 points indicating considerable difficulty learning narrative language, 1 point indicating some difficulty, and 2 points indicating very little difficulty. Two final scores were derived: (a) a total modifiability index of 14 possible points for the summed responses for Questions 1–7 (TMI), and (b) an overall modifiability rating from Question 7 alone with 0, 1, or 2 points possible (Mod-7). The 14-point TMI and 3-point Mod-7 scales were dichotomized post hoc to allow for the binary classification of LI/no LI. Examiners were blind to where the upper and lower points were collapsed to create binary scores.

Administration Training and Fidelity

Graduate and undergraduate students in speech-language pathology served as examiners. A Spanish-English bilingual, certified SLP (first author) trained six research assistants to collect language samples using the Frog Retell and to administer the DA. The four research assistants who administered the Spanish Frog Retells were sequential bilingual English-Spanish speakers.

For the Frog Retell training, research assistants collected five language samples from fellow research assistants in the language the research assistants would be using. For the DA, research assistants administered five practice sessions including using the modifiability rating form on fellow research assistants playing the role of children with and without LI. The first author observed and provided feedback until each examiner was able to deliver the DA procedures independently and accurately.

For administration fidelity, the first author observed 25% of all data collection procedures in the field to document fidelity of administration. Every examiner was observed at least once. For the Frog Retell, all the examiners followed the administration procedures, which included reading English and Spanish scripts from the SALT manual (J. F. Miller & Iglesias, 2012), with no omissions or deviations; this was further confirmed through review of the audio recordings. For the DA teaching phase, a fidelity checklist for the required steps and prompts showed 93% correct examiner execution, with a range of 88%–100%. As prescribed, the examiners taught the story grammar elements that the student omitted during the teaching phase. For the subordinating conjunctions, the examiners deviated from the instructions to only teach one conjunction, and instead taught any of the four on the score sheet that were missing. From the audio recordings after data collection was completed, the first author additionally reviewed a random selection of 32 (38%) of the 84 audio recordings. Most of the sessions were approximately 25 min in length and none exceeded 30 min. All but one session consisted of pretest, two to four full teaching cycles, and a posttest. For

Table 3. Prompt type and level used in the dynamic assessment teaching phase.

1. Correct and/or prompt immediately	Immediately stop the child if there is an error or an omission of a target feature. (e.g., *Wait, you forgot to tell me the problem*.)
2. Use least-to-most verbal prompting	Use a two-step prompting procedure: Level 1: Open-ended question (e.g., *What was the problem?*) Level 2: Model the target (e.g., *John crashed his bike and hurt his knee. Now you say that.*)
3. Use overcorrection procedure	Use the overcorrection procedure so that the child produces the target feature multiple times and has the opportunity to produce the target feature in context. Go back one step in the story before the target element so the child has an opportunity to produce the target in a meaningful context. (e.g., *That's right, John crashed his bike and hurt his knee. That's our problem; What was the problem?* [Child answers]. *Right! Almost every story has a problem. He crashed his bike and hurt his knee. What was the problem?* [Child answers]. *Now start here (point to preceding story grammar element) and keep going with the story. Remember to tell me the problem*.)
4. Foster independence	Especially during Steps 3 and 4 of the teaching phase, use the least amount of verbal prompting possible.

one participant, only one teaching cycle was completed in the time allowed.

Test Training and Reliability

Transcription and C-unit segmentation agreements of the Frog Retell language samples were completed following the data collection phase. The second author trained seven undergraduate students in communication disorders, three of whom were bilingual. The research assistants were blind to the participants' impairment status and the purposes of the study. After all of the Frog Retell language samples were transcribed and segmented, 30% were transcribed and segmented by research assistants who did not serve as the primary transcribers. Mean point-to-point agreement was 93% for transcription of words and 91% for C-unit segmentation. Coding was conducted using the SALT software (J. F. Miller & Iglesias, 2012); therefore, scoring agreement was unnecessary.

Examiners were trained to score the DA pretests and posttests in real time. The first author provided training and feedback until each examiner could score the DA pre- and posttests with 90% or greater point-to-point accuracy. This real-time scoring training took approximately 30 min for each examiner. During the data collection phase, the first author independently rescored 20% of the pre- and posttest retells in real time. Mean point-to-point agreement for real-time scoring of the pretests and posttests between the examiners and the first author was 92% (range 78%–100%).

To assess the reliability of the modifiability (TMI and Mod-7) ratings, the first author randomly selected and scored 16 (38%) of the first DA session recordings. Children with and without LI were present in the sessions examined. Interrater agreement for each of the first six modifiability items that went into TMI was 85%, and ranged from 14% to 100%, with a median and mode of 100%. The only interrater agreements below 80% were outliers of 14% and 43%. For the lowest agreement, the original rater assigned scores of 1 for a child with LI whenever the first author assigned scores of 0, except for the "student was easy to teach" category, which was assigned a 0 by both raters. For the 43% agreement, the original rater assigned scores of 1 for a TD child whenever the first author assigned scores of 2, except for the "high response to prompts," "easy to teach," and "displayed few disruptions," which both raters scored as 2.

For the Mod-7 item, mean point-to-point interrater agreement was 88%. Disagreements on Mod-7 were as follows: The original rater assigned a score of 1 and the first author assigned a score of 2 to a participant who was TD, and the original rater assigned a score of 0 and the first author assigned a score of 1 to a participant with LI.

As part of the exploratory investigation of whether the first 5–10-min teaching cycle could be substituted for a full teaching phase, the first author scored and compared modifiability ratings after the first and last teaching cycles for each of the 16 randomly selected audio-recorded sessions. Mean point-to-point intrarater agreement for each of the modifiability items was 97% (86% to 100%), with 13 of the 16 rating pairs in 100% agreement, and the remaining three pairs at 86% agreement, yielding a median and mode of 100%. Intrarater agreement specifically for Mod-7 after the first intervention cycle and the last intervention cycle was 100%.

To further establish the reliability of using only the first 5–10-min cycle of teaching, interrater reliability for this cycle was examined. To do this, the last author, a researcher and certified SLP, rated modifiability after the first treatment cycle of each session for the 16 randomly selected sessions used for intrarater reliability. Point-to-point interrater agreement for the first six modifiability items between the last author and the first author was 98%. For the overall modifiability score (Mod-7), the first and last authors agreed 100% of the time. Their ratings were also in perfect agreement with the "true" child classifications.

One of the data analyses involved setting optimized cut points for TD children versus children with LI rather than using the 3-point scale. For TMIs dichotomized between 12 and 11 points, point-to-point interrater agreement for the first 25-min DA session was 100%. When Mod-7 scores were dichotomized between 2 and 1, interrater agreement for the first 25-min DA session was 94%.

For the timing of duration for the first teaching cycle, interrater agreement was examined for a random selection of 5 (30%) of the 16 timing data points. The second independent timing to a precision of plus or minus 3 s by the last author resulted in 100% interrater agreement.

Results

This study investigated the classification accuracy of two 25-min sessions of a concentrated narrative DA for bilingual children with and without LI. DA indices from the first and both sessions were analyzed to determine the best predictors and most parsimonious combination of predictors. Those indices that were significantly different between children with and without LI were then submitted to step-wise discriminant function analyses to identify the variables that uniquely contributed to classification accuracy. ROC analyses were conducted to determine optimal cutoffs on the basis of sensitivity and specificity results.

Parsimonious Classification Accuracy of DA Indices

The first stage of the analysis examined group differences on DA gain, posttest scores, modifiability, and duration scores. To identify the variables most likely to be valuable predictors, independent *t* tests were conducted on the mean performance of the true impaired and true typical groups on the DA indices of: (a) Mod-7, (b) TMI, (c) posttest, (d) gain, and (e) duration (see Table 4). Independent *t* tests with an alpha level set to .01 to control for family-wise error showed significant differences between the children with and without LI on all the indices except the gain scores.

Step-wise discriminant function analysis was conducted to identify the variables that uniquely contributed

Table 4. Mean dynamic assessment performance of participants with language impairments (LI) and participants who are typically developing (TD).

	LI (*SD*)	TD (*SD*)	*t* value	*df*	*p* value	*d* value
TMI S1*	5.20 (3.49)	12.84 (1.64)	–9.59	40	< .001	2.80
TMI S2*	5.40 (2.68)	12.97 (1.60)	–11.19	40	< .001	3.43
Mod-7 S1*	0.50 (0.53)	1.88 (0.34)	–9.80	40	< .001	3.10
Mod-7 S2*	0.50 (0.53)	1.88 (0.34)	–9.80	40	< .001	3.10
Posttest S1*	8.90 (6.42)	15.94 (5.22)	–3.53	40	< .001	1.20
Posttest S2*	7.50 (5.91)	17.13 (6.15)	–4.36	40	< .001	1.60
Gain S1	4.90 (5.63)	4.56 (5.91)	0.16	40	= .87	0.09
Gain S2	-2.00 (4.76)	1.47 (5.41)	–0.87	40	= .39	0.68
Duration*	659 (178)	458 (156)	3.43		< .001	1.97

Note. TMI = total modifiability index; S1 = dynamic assessment Session 1; S2 = dynamic assessment Session 2; Mod-7 = overall modifiability score; Duration = duration of first teaching cycle in seconds.

*Significant at adjusted alpha of .01.

to classification accuracy for the first 25-min DA session and for both sessions. All the significant variables from the first 25-min DA session were entered in the model, and then at each step, the variable that contributed the least to the classification was discarded. For the first 25-min DA session, the overall Wilks's lambda was significant, $\lambda = .28$, $\chi^2(1, 42) = 44.00$, $p = < .001$. The analysis identified Mod-7 and duration as uniquely significant indicators of LI. These two variables combined yielded 90% sensitivity and 97% specificity.

The discriminant analysis was conducted again on the significant indices from both DA sessions (S1 and S2). This analysis maintained only Mod-7 from both sessions in the model, with duration discarded. The overall Wilks' lambda for these two overall modifiability variables was significant, $\lambda = .19$, $\chi^2(1, 42) = 58.84$, $p = < .001$, with 100% sensitivity and 100% specificity.

Establishment of Classification Cut Points

We then examined the combined classification accuracy of the significant indices from S1, and then again from S1 and S2 together, using ROC curve analysis using logistic regression predicted probability indices. Logistic regression uses binary scores (e.g., impaired vs. typical) as the dependent variable, and reports the probability of achieving a particular outcome for the dependent variable using the different predictor variables. A ROC curve plots the true positive rate (sensitivity) against the false-positive rate (1-specificity) for all possible cut points, providing information on classification accuracy and the respective cut points for each indicator. The area under the curve (AUC) ranges from .5 to 1.0. Models that provide no better than chance prediction have an AUC of .5, and models that are perfectly predictive have an AUC of 1.0.

Results of the ROC analyses for individual and combined indices are presented in Table 5. The AUC for the Mod-7 scores from S1 was .97, with 100% sensitivity and 88% specificity using a cutoff of 1 or lower. The combined Mod-7 scores from S1 and S2 resulted in an AUC of 1.00 (100% sensitivity, 100% specificity), with a cutoff of 1 or lower. AUC results for TMI for two sessions was strongest, with sensitivity of 100% and specificity of 94% using a cutoff of 11 or lower to indicated impairment. However, a cutoff of 10 for S1 still resulted in sensitivity and specificity over 90%. For posttest, the AUC for a cutoff of 14 or 15 was lower than for Mod-7 or TMI, with sensitivity ranging from 80% to 90% and specificity ranging from 72% to 91%. For duration, the AUC resulted in only 80% sensitivity and 81% specificity using a cutoff of 560 s (9 min, 20 s).

These results show that examiners could use a single indicator, preferably Mod-7, to identify children with and without LI with reasonable accuracy. However, examiners could also use a combination of DA indices to increase confidence in a judgment that a child has a LI. Student scores below the cut point for any two of the four indices (e.g., Mod-7 and Duration, Mod-7 from S1 and S2, or TMI plus posttest) resulted in 100% sensitivity and 91% specificity. However, accuracy decreased when more than two indicators were used. Performance below the cut point for any three out of the four indices resulted in sensitivity of 90% and specificity of 81%. Using all four indices resulted in only 70% sensitivity and 56% specificity.

Discussion

DA of narratives has a solid research base supporting its use for identifying LI and guiding treatment planning for children from culturally and linguistically different backgrounds. However, DA has not been adopted in the schools, due perhaps to any or all of several factors, including previously reported lengthy administration and scoring time, discomfort with or lack of information on the subjective aspect of modifiability ratings, or the lack of established score cut points to indicate typical development versus LI. This study investigated the classification accuracy of a more efficient version of narrative DA by examining the sensitivity and specificity of several DA indices, including a quantitative modifiability of DA teaching duration, and the establishment of clear cut points on the most predictive DA indices.

Classification Accuracy of a Concentrated DA on Narratives

In this investigation, a concentrated DA with real-time scoring resulted in very high sensitivity and specificity for bilingual kindergarten to third grade students. Mean scores on posttest, modifiability, and duration indices showed significant differences between children with and without LI. These three indices obtained very good to excellent sensitivity and specificity after only one 25-min DA session. The best predictor was a qualitative, binary modifiability rating, taken in any of three ways: (a) as a single indicator from the first DA session (100% sensitivity and

Table 5. Classification accuracy of participants with language impairment versus typical development by cut points.

Measure	Cut point	Sensitivity (n = 10)		Specificity (n = 32)		AUC
		%	n	%	n	
Mod-7 S1	≤1	100	10	88	28	.97
Mod-7 S1 + S2	≤1	100	10	100	32	1.00
TMI S1	≤10	90	9	91	29	.98
TMI S1	≤11	100	10	81	26	.98
TMI S1 + S2	≤11	100	10	94	30	.99
Posttest S1	≤14	80	8	72	23	.80
Posttest S1 + S2	≤15	90	9	75	24	.89
Posttest S1 + S2	≤14	80	8	91	29	.89
Duration C1	≥560	80	8	81	26	.82

Note. Mod-7 = overall modifiability score (only Question 7 on modifiability form); S1 = dynamic assessment Session 1; S2 = dynamic assessment Session 2; AUC = area under the curve predictive accuracy; TMI = total modifiability index; C1 = Cycle 1 of the dynamic assessment teaching session.

88% specificity); (b) from two DA sessions (100% sensitivity and 100% specificity), or (c) from the first DA session combined with teaching duration (90% sensitivity and 97% specificity).

These results are consistent with that of other DA studies, which have found that DA yields high classification accuracy and that modifiability is more consistently predictive than posttest scores or gains in test scores (e.g., Peña et al., 2006, 2014; Peña & Iglesias, 1992; Petersen & Gillam, 2015; Ukrainetz et al. 2000). A single rating of overall modifiability has been identified as the best single indicator, and accuracy has been higher for two teaching sessions (e.g., Peña & Iglesias, 1992; Ukrainetz et al., 2000). Although classification accuracy in the current study was best with information from two DA sessions, the result of the first teaching session was already much higher than the rates for most norm-referenced tests applied to children of linguistically and culturally diverse backgrounds (Figueroa & Newsome, 2006; Gandara, 2010; National Research Council, 2002). Given that eligibility decisions should be based on multiple performance measures, using this English language narrative DA along with appropriate norm-referenced tests or other measures should produce as high or even higher sensitivity and specificity rates for this challenging population in a reasonable amount of time.

The results of the analysis of multiple DA indices showed the cost-benefit decisions that must be made in choosing both assessment measures and performance standards. Sensitivity to identifying a child with a true LI (sometimes called a *hit*) and specificity in only identifying those with true impairments (avoiding *false positives*) often operate in opposition: by lowering the cutoff, fewer children are falsely identified as language impaired but more children with LI are missed, whereas raising the cutoff does the opposite.

Of all the individual indicators, the single rating of overall modifiability explained most of the unique variance and was clearly superior with its perfect classification accuracy for two sessions. Nevertheless, the decision on which indicator to use is more complicated: The single rating of overall modifiability (Mod-7) had higher sensitivity at 100%, but the sum of the seven modifiability items (TMI) had higher specificity at 91%. A caution is that this single rating, like for Peña and Iglesias (1992), was made following six prior ratings, so whether it would be as reliable if done alone is unknown. Combining indices can strengthen accuracy: Requiring scores below the cut points on any two of the four indices for one DA session were almost always classified correctly (100% sensitivity and 91% specificity). However, more information is not always better: Classification accuracy was 90% sensitivity and 81% specificity for three indices and 70% and 56% for four indices. More than two indices resulted in the indices working against each other to lower rather than increase classification accuracy.

Is it more of a concern to miss children with LI or to falsely identify TD children? Although neither option is acceptable, the rates of correct identification of these culturally and linguistically different children are higher than those shown with the most widely used measures of norm-referenced testing (Spaulding, Plante, & Farinella, 2006). Furthermore, these results are just for a single, short administration of DA. Coupled with other assessment measures in a full, valid, and reliable evaluation, results would be expected to be even higher.

The Feasibility of Real-Time Narrative Scoring

Narratives are well recognized as a valuable addition to a language evaluation. However, the time required to elicit, transcribe, and analyze narrative samples is often more than is available to a clinician. Narrative DAs similarly have a well-established research base, but their clinical use is problematic because of the time required for scoring and analysis. For the current investigation, a real-time scoring and goal-setting procedure was used that had been developed in prior investigations of narrative language (Petersen & Spencer, 2012). This very efficient format

could easily be included in a typical language evaluation timeframe.

In this study, narrative retells were scored for presence of story grammar elements and adverbial conjunctions in real time with the aid of a structured scoring sheet. Interrater reliability was strong. An additional investigation was conducted on the student examiners listening to the audio recordings of the narrative retells one additional time (Pettipiece & Petersen, 2013). Those scores then showed 90%–95% agreement with the more experienced first author's real-time scores. Pettipiece and Petersen (2013) found that the extra listening took less than 2 min, but listening to the narrative again is still an extra step for an already highly reliable procedure. Furthermore, to use a second listening procedure to set teaching goals, this extra listening would have to occur during the session, resulting in a brief pause after the pretest while the examiner reviews the audio recording. On the other hand, the DA teaching phase may not require individualized goal selection, which would allow the extra listening to occur after the session. Examination of the audio records revealed that during the DA teaching phase, the examiner helped each child produce all of the story grammar elements and all of the adverbial subordinate conjunctions missing from their pretest retells. This resulted in typically four or five story grammar elements plus all four subordinating conjunctions with some participants. The advantage to this approach was the ease in which examiners could understand and implement this simple rule: Target any story grammar elements and subordinate conjunctions that the child does not produce in the pretest retell.

Lack of Differential Gain With This DA

This study investigated the contributions to classification accuracy of narrative posttest scores, narrative gain scores, two modifiability ratings, and teaching duration. The results of the initial comparative analysis determined that only gain scores showed no significant differences between the children with LI and those with typical development.

This lack of predictiveness of gain scores would seem to conflict with the conceptual base of DA (Gutiérrez-Clellen & Peña, 2001; Lidz & Peña, 1996; Vygotsky, 1978). A child with an intact language-learning mechanism and thus a better potential for learning should show greater progress from a brief teaching session than a child with an impairment. This difference should be revealed in differential gains from pretest to posttest. However, in the current study, impaired learners gained as much, on average, as typical learners. Peña et al. (2014) noted similar findings. They reported no differential gains on story grammar, language productivity measures, or grammaticality between two groups of TD children and children with LI. In addition to consistently lacking discriminatory power, psychometric problems can result from using gain scores (Elliott, 2003; Embretson, 1987; Sternberg & Grigorenko, 2002).

One possible reason for the lack of differential gains is that the teaching primarily addressed narrative structure, which is relatively easy to learn. In one short session, it is reasonable to expect that even a child with LI might learn to provide a motivation, an attempt, or a resolution to a problem. Syntactic structures are generally more difficult to teach, but the subordination addressed involved four particular conjunctions that are relatively easy to highlight. In addition, posttest retells were done immediately, on different stories but using the same elicitation format, showing immediate retention but not longer term use. Last, the teaching procedure was explicit and systematic, using maximal support by repeatedly cycling through a retell with models, prompts, and visual supports. Nevertheless, despite all these ways that the learning should have been easy, the two types of children did not show ceiling effects: Even the mean scores for the TD children were only about half of possible, leaving that critical "room to improve."

Thus, despite the lack of differential gains, the purpose of assessment was accomplished via the other indicator type: modifiability ratings. In this DA, extra effort was expended by both child and examiner when disability was present. This difference in effort was revealed in both the sum of the modifiability items and the single overall rating of modifiability, consistent with other DA studies (e.g., Peña & Iglesias, 1992; Ukrainetz et al., 2000). This DA format, despite its brevity and ease of administration, sufficiently stressed the systems of the children with LI so that they exhibited greater levels of frustration, disruptions, and inattentiveness—and the examiners had to work harder to get them to their goals.

Concentrated DA: Can It Be Further Concentrated?

A major feature of this investigation was condensing the DA process. Prior DA research has entailed three to four separate testing and teaching sessions. DA using narratives has also necessitated additional time for transcription and scoring. Instead, in this study, the DA format investigated consisted of test–teach–test plus scoring concentrated into a single session of less than half an hour. The testing phases each took less than 3 min and the teaching phase took 15–20 min. In that time, all but one session consisted of pretest, two to four full teaching cycles, and a posttest in 25 min. One participant was able to complete only one teaching cycle the maximum time of 30 min.

Although this DA session is considerably shorter than what has been used in previous research, it may be possible to abbreviate the process even more. The first teaching cycle in the first DA session was almost always completed within 10 min. If the teaching phase could be reduced to just one brief cycle and still result in acceptable classification accuracy, then DA could easily be used within a typical language evaluation.

The duration indicator, which involved timing the first teaching cycle of the first DA session, certainly showed promise. This simple quantitative indicator of response to teaching explained a significant amount of unique variance in the discriminant function analysis, more than posttest or gain scores.

Because of the promising duration results, the investigators took a closer post hoc look at the first teaching cycle. Examination of the intrarater agreements used for reliability revealed that the modifiability ratings (Mod-7 and TMI), when scored after the first teaching cycle, were nearly identical to the same ratings calculated at the end of the 25-min session. This degree of consistency in modifiability ratings has been documented in other research. For example, Peña et al. (2007) found a very high level of consistency in scoring modifiability across two teaching sessions ($r > .95$). These findings suggest that the excellent classification accuracy obtained from a single 25-min DA session could possibly be obtained in less than 15 min.

The question could be raised about eliminating the testing phases. Some versions of DA are teaching only (e.g., Feuerstein, 1980; Olswang & Bain, 1996; Olswang, Feuerstein, Pinder, & Dowden, 2013; Stock-Shumway, 1999). However, in this DA, the pretest was needed to provide the initial narrative model for the teaching phase. Although the posttest scores were less helpful than the other indicators in identifying LI, they are objective, quantitative measures that are likely to be more accepted clinically. In addition, the posttest informs subsequent language intervention by both immediate intervention targets in terms of those that were temporarily achieved within DA, and longer term intervention targets of those that were still missing at DA posttest.

These results suggest that DA may be reducible to a single very short test–teach–test session, using two indicators: (a) a simple, objective, reliable measure of teaching duration, and (b) a single rating of overall modifiability. This highly concentrated DA merits further investigation.

Limitations and Future Research

The findings of this study suggest that narrative DA can be administered and scored with efficiency and high classification accuracy. This DA procedure shows strong potential for clinical adoption, but several aspects of this need to be investigated further before deploying it clinically.

Duration of teaching and modifiability ratings obtained after only one teaching cycle suggest that DA may be reducible even further, but planned investigations are required for these preliminary findings. Exploring the temporal limits of DA would contribute to improving the appeal of the procedure for clinical application. Another research need is to apply the indices, scoring procedures, and cut points to a new sample of learners to determine if these statistically optimized and sample-specific results continue to yield this high classification accuracy.

A next step for determining clinical viability is determining whether the reliability of the modifiability ratings obtained with trained research assistants would be similarly high in a clinical setting. The training sessions were only 30 min long, but that is certainly more than is typically required to learn to score a norm-referenced test. The amount of training that would be required for experienced clinicians to score the narrative DA in real time needs to be established. Another line of inquiry is further investigation of the viability of the single overall rating of modifiability. Being able to judge learner potential on the basis of a single question, which was obtained in this and other studies (e.g., Peña & Iglesias, 1992), is appealing but needs strong evidence. The predictivenss of a single rating without being proceeded by a half-dozen focused ratings needs to be determined. Furthermore, the examiners were unaware of the cut points for dichotomizing the modifiability scales. It is unclear to what extent a priori knowledge of cut points for LI will influence examiner's judgments of modifiability.

This study investigated the accuracy of this DA procedure with bilingual Spanish-English early elementary–grade children. This means that the study has the standard limitation that the findings cannot be unquestionably generalized to students from all other cultural and linguistic groups. Furthermore, the sample size of 42 was small, with a small base rate and some (although statistically nonsignificant) differences between the groups of children with and without LI, so the results should be considered with caution. Although this study lacked sufficient power to examine differential effects of the DA on children with divergent language profiles, future research should explore this possibility. Also, the diagnosis of LI that was the classification reference for this study involved use of measures that are known to be biased and problematic for this population. To surmount this, multiple evaluation sources were used, including Spanish language samples, teacher and parent input, and involvement of a bilingual SLP. However, some of the children could have been initially inaccurately identified as "true language impaired" or "true typically developing." Researchers should explore the possibility of collecting longitudinal data on student language learning as a gold-standard indicator as to whether a student truly has LI.

The evidence from this study and others is clear on the value of DA in language evaluation for children from culturally and linguistically different backgrounds. An important next step in this line of research is moving from a research context to clinical implementation of this efficient, accurate assessment procedure (Olswang & Prelock, 2015). SLPs should be invited to use and evaluate this version of narrative DA in their regular work settings. Routes and roadblocks to regular use of DA can then be identified, and further refinements of this valuable procedure can be investigated.

As part of increasing implementation of DA, SLPs and other educators should be guided to understand the fundamental similarities—and benefits—of DA compared to the widely used *response to intervention* (RTI, or its newer label of *multitiered systems of support*). Federal law permits RTI to be used to identify specific learning disability (CFR 300.307, from http://ecfr.gpoaccess.gov/) in addition to its primary use as a noncategorical instructional support framework to improve at-risk children's reading performance. Best practices in RTI are still emerging; however, it has shown clear benefits over the traditional discrepancy method for identifying specific reading disability (e.g., Fuchs, Fuchs, & Compton, 2004; Swanson &

Howard, 2005). RTI and DA share the conceptual base of using a child's response to teaching to determine learning potential. However, DA may be more helpful than RTI for this purpose because it provides a "purer" estimate of learning potential. RTI is typically delivered to large numbers of at-risk students as a set curriculum by a variety of educators over a whole school year. This means that RTI is likely more susceptible to experiential factors, such as child attendance, peer interactions, teaching quality, and cultural-linguistic mismatches than DA with its short, uniform, individually administered procedures. DA, in contrast to RTI (or multitiered systems of support) has as its primary purpose, not to make lasting change, but rather to determine learning potential. Having SLPs and other educators understand that DA is fundamentally a miniature version of RTI, and one that is perhaps better suited to diagnostic decisions, might contribute to it being more broadly accepted in clinical practice.

Conclusion

This study shows that narrative DA can be delivered efficiently and accurately in a concentrated test–teach–test format that provides both assessment and treatment information. Modifiability ratings, teaching duration, and posttest scores from one or two 25-min DA sessions, administered in English, resulted in excellent classification of bilingual kindergarten to third graders with and without LI. Preliminary results suggest that this short DA has the potential to be further abbreviated and still retain its classification accuracy. Empirically based cut points were established to allow for clinical use for classification decisions. This narrative DA format shows strong potential to be combined with conventional assessment measures to improve the identification of LI in children from culturally and linguistic diverse backgrounds.

Acknowledgments

This study was funded by the Barbara Kahn Foundation for the Division of Communication Disorders, University of Wyoming. The authors thank Michelle Buchanan for her meaningful contributions, and Brenna Thompsen, Penny Tonn, Gwynn Barrow, Mike Soriano, Erin Bayliff, Mckenzie Dickerson, Haley Perl, Shaylyn Sheaffer, Brianna Schwan, and Anya Tracy for their help in data collection and analysis.

References

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44,* 133–146.

Bishop, D. V. M., & Edmundson, A. (1987). Language impaired 4-year-olds: Transient from persistent impairment. *Journal of Speech and Hearing Disorders, 52,* 156–173.

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.

Ebert, K. D., & Scott, C. M. (2014). Relationships between narrative language samples and norm-referenced test scores in language assessments of school-age children. *Language, Speech, and Hearing Services in Schools, 45,* 337–350.

Elliott, J. (2003). Dynamic assessment in educational settings: Realising potential. *Educational Review, 55,* 15–32.

Embretson, S. E. (1987). Towards development of a psychometric approach. In C. S. Lidz (Ed.), *Dynamic assessment: An interactional approach to evaluating learning potential* (pp. 141–170). New York: Guilford Press.

Fazio, B. B., Naremore, R. C., & Connell, P. J. (1996). Tracking children from poverty at risk for specific language impairment: A three-year longitudinal study. *Journal of Speech and Hearing Research, 39,* 611–624.

Feuerstein, R. (1980). *Instrumental enrichment: An intervention program for cognitive modifiability*. Baltimore, MD: University Park Press.

Figueroa, R. A., & Newsome, P. (2006). The diagnosis of LD in English learners: Is it nondiscriminatory? *Journal of Learning Disabilities, 39,* 206–214.

Fuchs, D., Fuchs, L. S., & Compton, D. L. (2004). Identifying reading disabilities by responsiveness-to-instruction: Specifying measures and criteria. *Learning Disability Quarterly, 27,* 216–227.

Gandara, P. (2010). The Latino education crisis. *Educational Leadership, 67*(5), 24–30.

Gipps, C. (1999). Socio-cultural aspects of assessment. *Review of Research in Education, 24*(1), 355–392.

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

Hasson, N., Camilleri, B., Jones, C., Smith, J., & Dodd, B. (2012). Discriminating disorder from difference using dynamic assessment with bilingual children. *Child Language Teaching & Therapy, 29,* 57–75.

Hasson, N., Dodd, B., & Botting, N. (2012). Dynamic Assessment of Sentence Structure (DASS): Design and evaluation of a novel procedure for assessment of syntax in children with language impairments. *International Journal of Language & Communication Disorders 47,* 285–299.

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

Haynes, W., & Pindzola, R. (2007). *Diagnosis and evaluation in speech pathology* (8th ed.). Boston, MA: Allyn & Bacon.

Kramer, K., Mallett, P., Schneider, P., & Hayward, D. (2009). Dynamic assessment of narratives with grade 3 children in a First Nations community evaluation. *Canadian Journal of Speech-Language Pathology & 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.

Lidz, C., & Peña, E. (1996). Dynamic assessment: The model, its relevance as a nonbiased approach, and its application to Latino American preschool children. *Language, Speech, & Hearing Services in Schools, 27,* 367–372.

Lopez, E. C. (1997). The cognitive assessment of limited English proficient and bilingual children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues.* New York, NY: Guilford Press.

Mayer, M. (1969). *Frog, where are you?* New York, NY: Dial Press.

McFadden, T. U. (1996). Creating language impairments in typically achieving children: The pitfalls of "normal" normative

sampling. *Language, Speech, and Hearing Services in Schools, 27,* 3–9.

Miller, J. F., & Iglesias, A. (2012). *Systematic analysis of language transcripts* (Research Version 2012.4.5) [Computer software]. Middleton, WI: SALT Software LLC.

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

National Research Council. (2002). *Minority students in special and gifted education*. Washington, DC: National Academy Press.

Olswang, L. B., & Bain, B. A. (1996). Assessment information for predicting upcoming change in language production. *Journal of Speech, Language, and Hearing Research, 39,* 414–423.

Olswang, L. B., Feuerstein, J. L., Pinder, G. L., & Dowden, P. (2013). Validating dynamic assessment of triadic gaze for young children with severe disabilities. *American Journal of Speech-Language Pathology, 22,* 449–462.

Olswang, L. B., & Prelock, P. A. (2015). Bridging the gap between research and practice: Implementation science. *Journal of Speech, Language, and Hearing Research, 58,* S1818–S1826.

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairments in English language learners. *Journal of Speech, Language, and Hearing Research, 57,* 2208–2220.

Peña, E. D., Gillam, R. B., Malek, M., Ruiz-Felter, R., Resendiz, M., Fiestas, C., & Sabel, T. (2006). Dynamic assessment of school-age children's narrative ability: An experimental investigation of classification accuracy. *Journal of Speech, Language, and Hearing Research, 49,* 1037–1057.

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

Peña, E., Resendiz, M., & Gillam, R. (2007). The role of clinical judgments of modifiability in the diagnosis of language impairment. *Advances in Speech Language Pathology, 9,* 332–345. https://doi.org/10.1080/14417040701413738

Petersen, D. B., Allen, M. M., & Spencer, T. D. (2016). Predicting reading difficulty in first grade using dynamic assessment of decoding in early kindergarten: A large-scale longitudinal study. *Journal of Learning Disabilities, 49,* 200–215.

Petersen, D. B., Brown, C., Ukrainetz, T. A., Wise, C., Spencer, T. D., & Zebre, J. (2014). Systematic individualized narrative language intervention on the personal narratives of children with autism. *Language, Speech, and Hearing Services in Schools, 45,* 67–86.

Petersen, D. B., & Gillam, R. B. (2013). Accurately predicting future reading difficulty for bilingual Latino children at risk for language impairment. *Learning Disabilities Research & Practice, 28,* 113–128.

Petersen, D. B., & Gillam, R. B. (2015). Predicting reading ability for bilingual Latino children using dynamic assessment. *Journal of Learning Disabilities, 48,* 3–21.

Petersen, D. B., Gillam, S. L., Spencer, T. D., & Gillam, R. B. (2010). The effects of literate narrative intervention on children with neurologically based language impairments: An early stage study. *Journal of Speech, Language, Hearing Research, 53,* 961–981.

Petersen, D. B., & Spencer, T. D. (2010). *Test of Narrative Retell: School Age*. Laramie, WY: Language Dynamics Group. Retrieved from http://www.languagedynamicsgroup.com/

Petersen, D. B., & Spencer, T. D. (2012). The narrative language measures: Tools for language screening, progress monitoring, and intervention planning. *Perspectives on Language Learning and Education, 19,* 119.

Petersen, D. B., & Spencer, T. D. (2014). *The Predictive Early Assessment of Reading and Language (PEARL)*. Palmer, AK: Language Dynamics Group. Retrieved from http://www.languagedynamicsgroup.com/

Petersen, D. B., Thompsen, B., Guiberson, M., & Spencer, T. D. (2015). Cross-linguistic interactions from second language to first language as the result of individualized narrative language intervention with children with and without language impairment. *Applied Psycholinguistics, 37,* 703–724.

Pettipiece, J., & Petersen, D. B. (2013, November). *Interrater reliability of real-time, recorded audio, and transcribed scoring of school-age bilingual children's narratives*. Poster presented at the annual American Speech-Language-Hearing Association Conference, Chicago, IL.

Spaulding, T. J., Plante, E., & Farinella, K. A. (2006). Eligibility criteria for language impairment: Is the low end of normal always appropriate? *Language, Speech, and Hearing Services in Schools, 37,* 61–72.

Spaulding, T. J., Szulga, M. S., & Figueroa, C. (2012). Using norm-referenced tests to determine severity of language impairment in children: Disconnect between U.S. policy makers and test developers. *Language, Speech, and Hearing Services in Schools, 43,* 176–190.

Spencer, T. D., Kajian, M., Petersen, D. B., & Bilyk, N. (2014). Effects of an individualized narrative intervention on children's storytelling and comprehension skills. *Journal of Early Intervention, 35,* 243–269.

Spencer, T. D., & Petersen, D. B. (2012). *Story Champs*. Palmer, AK: Language Dynamics Group. Retrieved from http://www.languagedynamicsgroup.com/

Spencer, T. D., Petersen, D. B., Slocum, T. A., & Allen, M. M. (2014). Large group narrative intervention in Head Start preschools: Implications for response to intervention. *Journal of Early Childhood Research, 13,* 196–217.

Spencer, T. D., & Slocum, T. A. (2010). The effect of a narrative intervention on story retelling and personal story generation skills of preschoolers with risk factors and narrative language delays. *Journal of Early Intervention, 32,* 178–199.

Sternberg, R. J., & Grigorenko, E. L. (2002). *Dynamic testing: The nature and measurement of learning potential*. New York, NY: Cambridge University Press.

Stock-Shumway, K. (1999). *The effectiveness of dynamic assessment as a kindergarten screening measure* (unpublished master's thesis). University of Wyoming, Laramie.

Swanson, H. L., & Howard, C. B. (2005). Children with reading disabilities: Does dynamic assessment help in the classification? *Learning Disability Quarterly, 28,* 17–34.

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.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, MA: Harvard University Press.

Weddle, S., Spencer, T. D., Kajian, M., & Petersen, D. B. (2016). An examination of a multitiered system of language support for culturally and linguistically diverse preschoolers: Implications for early and accurate identification. *School Psychology Review, 45,* 109–132.

Wetherell, D., Botting, N., & Conti-Ramsden, G. (2007). Narrative in adolescent specific language impairment (SLI): A comparison with peers across two different narrative genres. *International Journal of Language & Communication Disorders, 42,* 583–605.

Appendix A

Sample Story for DA Testing and Teaching

Character/setting	Problem	Feeling	Plan/attempt	Consequence/ end emotion
One day, John was riding his bike down a rocky street because he wanted to go to a friend's house that was far away.	But John accidentally crashed into a rock and cut his knee.	John was sad because the cut hurt.	When he got up he decided to get help at home. John quickly ran home and said to his mom "I need a Band-Aid."	Then his mom said "I have just what you need." She put a big blue Band-Aid on his cut. After John got the Band-Aid, John's knee felt better. Then he was happy because he could go back outside to ride his bike.

Note. Story from *Story Champs* language intervention (Spencer & Petersen, 2012).

Appendix B

Modifiability Rating Form

		2	1	0
1	Student had high response to prompts			
2	Student exhibited high degree of transfer			
3	Student attended to the teaching			
4	Student was easy to teach			
5	Student displayed limited frustration			
6	Student displayed few disruptions			
7	What is your overall judgment of the student's potential to learn narrative language? (MOD 7)			
Total Modifiability Index (TMI)				

1. Response to Prompts

2 points = Examiner provides prompt and student responds appropriately most of the time. Little redirection required. Prompts are more Level 1 (open-ended questions) than Level 2 (examiner models). Student quickly retells elements without examiner telling student what to say.

1 point = Examiner provides prompt and student responds appropriately some of the time. Some redirection required. Requires more Level 2 prompts than Level 1 prompts for student to respond correctly.

0 points = Examiner provides prompt and student responds appropriately infrequently. Considerable redirection required. Almost all Level 2 prompts (examiner models). Student pauses a long time.

2. Degree of Transfer

2 points = Transfer of one or two targets is evidenced often as student progresses within and across cycles. One or two targets are frequently transferred across cycles (from one story to the next). All story grammar elements with interchangeable plan/attempt, and the word *because* is produced in last step of last cycle (no pictures/no icons) with no more than one Level 1 prompt.

1 point = Transfer of one target is evidenced occasionally as student progresses within and across cycles. One target is occasionally transferred across cycles (from one story to the next). Many story grammar elements with interchangeable plan/attempt, and the word *because* is produced in last step of last cycle (no pictures/no icons) with two or more Level 1 prompts.

0 points = Transfer of targets is evidenced rarely as student progresses within and across cycles. Targets are rarely transferred across cycles (from one story to the next). Some story grammar elements (five or fewer) are produced in last step of last cycle (no pictures/no icons) with one or more Level 2 prompts.

3. Attention to Teaching

2 points = On task. No verbal redirects to attend. Completely understands tasks. Attentive and focused.

1 point = Student is on task some of the time. Examiner is required to redirect attention some of the time. Student understands tasks some of the time. Distractible, but can be refocused.

0 points = Student often does not understand tasks (<25% of time). Examiner required to redirect attention much of the time. Understands tasks some of the time. Distracted and difficult to refocus.

4. Ease of Teaching

2 points = Minimal effort required to induce change. Effort greatly decreases within and across cycles. Examiner has to start few or no principles or examples.

1 point = Some effort required to change. Effort decreases somewhat within and across cycles. Examiner has to state some principles or examples.

0 points = Considerable effort required to induce change. Effort decreases very little within and across cycles. Examiner has to state many principles or examples.

5. Frustration

2 points = Verbal and nonverbal behavior that indicates little or no frustration. Appears to be happy with responses. Smiles or says, "I like this," or "This is easy." Calm, little to no soothing required. Enthusiastic, engages in tasks readily. Persistent. Wants to continue despite difficulty.

1 point = Verbal and nonverbal behavior that indicates some frustration. Uncomfortable. Breaks needed to soothe. Ambivalent. Unsure about tasks. Tentative. Appears unsure about continuing.

0 points = Verbal and nonverbal behavior that indicates considerable frustration. When students looks to examiner for help or says, "I don't know" or "I can't." Distressed. Much soothing required.

6. Disruptions

2 points = Little to no verbal and nonverbal behavior that interrupts flow of the intervention. Cooperative. Does not shift in seat.

1 point = Some verbal and nonverbal behavior that interrupts flow of the intervention. Changes topic once or twice. Shifts in seat on occasion. Looks at something other than the stimulus items once or twice.

0 points = Considerable verbal and nonverbal behavior that interrupts flow of intervention. Uncooperative. Refusing. Frequently changes topic. Frequently shifts in seat. Looks at something other than the stimulus book often.

7. Overall Judgment of Student's Potential to Learn Narrative Language

On a scale of 0–2, rate the student's potential to learn narrative language based on your interaction with the student during the teaching phase of the dynamic assessment.