AJSLP

Research Article

Dynamic Assessment Language Tasks and the Prediction of Performance on Year-End Language Skills in Preschool Dual Language Learners

Janet L. Patterson,[a] Barbara L. Rodríguez,[a] and Philip S. Dale[a]

Purpose: Early identification is a key element for accessing appropriate services for preschool children with language impairment. However, there is a high risk of misidentifying typically developing dual language learners as having language impairment if inappropriate tools designed for monolingual children are used. In this study of children with bilingual exposure, we explored performance on brief dynamic assessment (DA) language tasks using graduated prompting because this approach has potential applications for screening. We asked if children's performance on DA language tasks earlier in the year was related to their performance on a year-end language achievement measure.
Method: Twenty 4-year-old children from Spanish-speaking homes attending Head Start preschools in the southwestern United States completed three DA graduated prompting language tasks 3–6 months prior to the Head Start preschools' year-end achievement testing. The DA tasks, Novel Adjective Learning, Similarities in Function, and Prediction, were administered in Spanish, but correct responses in English or Spanish were accepted. The year-end achievement measure, the Learning Accomplishment Profile–Third Edition (LAP3), was administered by the children's Head Start teachers, who also credited correct responses in either language.
Results: Children's performance on two of the three DA language tasks was significantly and positively related to year-end LAP3 language scores, and there was a moderate and significant relationship for one of the DA tasks, even when controlling for age and initial LAP3 scores.
Conclusions: Although the relationship of performance on DA with year-end performance varies across tasks, the findings indicate potential for using a graduated prompting approach to language screening with young dual language learners. Further research is needed to select the best tasks for administration in a graduated prompting framework and determine accuracy of identification of language impairment.

Accurate early identification of language impairment is important for access to appropriate services. Children with bilingual experience are at risk for misidentification if cultural and linguistic diversity is not adequately addressed to differentiate disorders from differences in screening and assessment practices. Dual language learners (DLLs) are young children from homes in which at least one parent speaks a language other than English and who are learning two languages at the same time (Paradis et al., 2011). DLLs presently comprise nearly one third of all young children in the United States and more than 20% of young DLLs in 24 states and the District of Columbia (Park et al., 2017). Since 2000, there has been a 24% increase in the young U.S. DLL population. A closer examination of the young DLL population reveals that 62% are Hispanic/Latino, 23% are 3–4 years of age, and 58% live in low-income households. There is linguistic diversity within this population, but a majority (59%) of parents of DLLs in the United States speak Spanish (U.S. Census Bureau, 2015).

DLLs may initially be underidentified as having disabilities if all variability in children's performance is attributed to learning a second language and/or cultural variability; conversely, language impairments and other disabilities may be overidentified if cultural and linguistic differences are not taken into account. It has been suggested that language impairment and other disabilities may be underidentified in young DLLs but may be overidentified in

[a]Department of Speech and Hearing Sciences, The University of New Mexico, Albuquerque

Correspondence to Janet L. Patterson: jpatters@unm.edu

Editor-in-Chief: Julie Barkmeier-Kraemer
Editor: Li Sheng

Received April 26, 2019
Revision received July 7, 2019
Accepted November 1, 2019
https://doi.org/10.1044/2019_AJSLP-19-00120
Publisher Note: This article is part of the Forum: Innovations in Clinical Practice for Dual Language Learners, Part 2.

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

older DLLs, as clinicians and educators begin attributing children's lower achievement to factors other than limited English proficiency (Hibel & Jasper, 2012; Samson & Lesaux, 2009).

Cultural and Linguistic Diversity Among DLLs

Children who are DLLs are a highly varied group from diverse family backgrounds with a variety of life experiences. Children's development cannot be understood isolated from the social, cultural, and historical contexts in which it occurs (Vygotsky, 1978). From a sociocultural point of view, children approach developmental tasks, including language development, embedded in environments imbued with cultural norms and familiar cultural practices (Rogoff, 2003). This perspective is particularly relevant for understanding the development of DLLs because their experiences differ in many ways from those of young monolingual children. The differences in children's social, economic, and cultural contexts of experience play a significant role in their performance on unfamiliar tasks, which are commonly a part of speech-language pathologists' language assessment tools and procedures (Henderson et al., 2018).

Among young DLLs from homes in which Spanish is spoken, there also is great variability in their specific experiences with Spanish and English. Some come from monolingual Spanish homes, while others are from homes in which Spanish and English are spoken; some attend monolingual English preschool programs, and some attend dual language programs (Hammer & Rodríguez, 2012). The relationships of children's experiences and their language development in each language vary, depending on whether the child's experiences with input (exposure) to each language and/or output (use of each language) are examined (Bohman et al., 2010). Relationships between experience and language development also vary depending on the aspects of language development examined, including language domain (semantic, morphosyntactic, phonological; Bohman et al., 2010; Hammer et al., 2012; Scarpino et al., 2018), receptive versus expressive modality (Ribot et al., 2018), and home versus school language development among children attending monolingual English schools (Gutierrez-Clellen & Kreiter, 2003; Hammer et al., 2012). The extensive diversity of linguistic experiences among bilingual children poses a challenge in designing language assessment and screening tools (Pearson, 1998).

Language Assessment and DLLs

There have been important gains in developing approaches and tools for more accurately identifying language impairment among bilingual children in recent years. These gains have addressed sociocultural diversity in experiences with assessment tasks through the use of dynamic assessment (DA) and linguistic diversity by considering a child's performance in both languages.

DA

DA has been widely recommended to address differences in sociocultural experiences. The focus of DA is on a child's modifiability or learning in response to support, instruction, and feedback, which are provided as needed during the assessment. However, specific approaches to DA, such as mediated learning experience (MLE) and graduated prompting methods, vary in structure and format (Kapantzoglou et al., 2012). In MLE approaches to DA, supports are provided on an individualized basis, with the assessor providing scaffolding to transform and organize the stimuli, environment, and information as needed for the learner. In contrast, "graduated prompting" approaches employ a hierarchy of scripted prompts that provide progressively more explicit information. Although mediated learning approaches yield more comprehensive information about young children's potential and emerging skills (Tzuriel, 2000), graduated prompting can provide some information about children's responses to support with greater efficiency (Poehner & Lantolf, 2013).

A variety of studies using DA procedures have shown moderate to excellent sensitivity and specificity in diagnosing language impairment among minority and bilingual children (e.g., Kapantzoglou et al., 2012; Peña et al., 2014, 2001; Petersen et al., 2017). These studies have used MLE approaches with pretest and posttest measures and have required pretesting and one or more sessions of about 30 min each.

In contrast with the use of DA in diagnostic assessment, there has been limited research on DA procedures to screen for language impairment in young DLLs. This may be partly due to the amount of time required for many dynamic language assessment procedures, particularly those using MLE approaches and pretest–teach–posttest formats. Graduated prompting holds promise for using DA in screening because it may take less time than MLE approaches and the scripting may be sufficient to allow administration by paraprofessionals. In addition, pretesting and posttesting are not required for graduated prompting approaches. With graduated prompting, scoring can reflect two important features of support during a single teaching session: (a) the number of prompts presented and (b) the extent of information included in each prompt.

There is some evidence that graduated prompting approaches to DA are predictive of future progress and difficulties. Kindergarten children's risk for reading disabilities was more accurately predicted by a phonological awareness task administered within a graduated prompting framework than the original, static version (Bridges & Catts, 2011). Among second-grade children, those with learning disabilities required two to three times more prompts on verbal and spatial reasoning tasks than children without learning disabilities (Resing, 1997, as cited by Tzuriel, 2000). In a study of a diverse group of bilingual (DLL) and monolingual English-speaking preschoolers referred for speech-language services, children's performance on a static standardized vocabulary measure was predictive of their performance on the measure 6 months later, but responses to a word-learning task with graduated prompting improved prediction, particularly for children with scores in the lowest quartile (Camilleri & Law, 2014). The contribution of DA to predicting future performance, particularly among children who score low on static measures, is of special interest for screening purposes as it indicates DA may be particularly helpful for reducing potential overidentification.

Although the children in the Camilleri and Law (2014) study were all from homes in which English was spoken, some were from homes in which a language other than English was also spoken. Camilleri et al. (2014) did an exploratory analysis of DA findings for 12 bilingual preschool children referred for speech-language therapy. The children were from homes with diverse languages, but the children all had at least some exposure to English, based on the study criterion of having attended English-speaking preschools for several months. One of the 12 children did well on all the DA tasks, which included vocabulary, syntax, and phonology, indicating the child most likely did not have a speech-language impairment. The authors interpret their findings as showing some potential for DA for "prediagnostic" purposes. Although limited and conducted only in English, these studies taken together indicate promise for use of DA for screening purposes with bilingual preschool children.

Language Assessment and Bilingual Considerations

Two assessment practices that are essential for addressing the linguistic experiences of young DLLs are (a) using tools that are appropriate for each language and (b) crediting children with their knowledge in both languages. In recent years, some assessment and screening tests have incorporated these features.

The Spanish version of the Preschool Language Scale–Fifth Edition (PLS-5 Spanish) and the PLS-5 Spanish Screening Test (PLS-5 SST) target Spanish grammatical targets, rather than being a direct translation of the English PLS-5. The Bilingual English–Spanish Assessment (Peña et al., 2016) and the Bilingual English–Spanish Oral Screening (BESOS; Lugo-Neris et al., 2015) address language-specific targets in both languages. An alternate approach to target selection that has been used to address dialectal variation and has potential for application with bilingual children was used for the Diagnostic Evaluation of Language Variation (Seymour et al., 2005), a norm-referenced test that was developed to reduce overidentification of African American English–speaking children due to dialectal variation. The Diagnostic Evaluation of Language Variation was constructed using items and skills that are noncontrastive for African American English and "mainstream American English"; Southern and Spanish-influenced varieties of English were also considered in the development of items and try-out testing (Seymour et al., 2005).

Approaches to scoring that credit bilingual children with knowledge in both languages also vary. A method that has been used frequently for expressive vocabulary and some other semantic measures is conceptual scoring. With conceptual scoring, a child's correct response in either language is credited but is not credited twice for correct responses in both languages for the same item or concept. Conceptual vocabulary scores for preschool and early school-age bilingual children are, of course, higher than single language scores, but they vary depending on the method of administration and scoring, and they are not as high as monolingual children's scores (Anaya et al., 2018). On the Bilingual English–Spanish Assessment (Peña et al., 2016) and the BESOS (Lugo-Neris et al., 2015), testing is done in each language separately, although correct responses in the other language are credited where appropriate. Scoring cutoffs are based on consideration of children's scores in each language. On the Spanish version of the PLS-5, the test is administered in Spanish, but responses in either Spanish or English are accepted (Zimmerman et al., 2015).

Both the PLS-5 SST and the BESOS have acceptable to high sensitivity; the PLS-5 SST correctly identifies 85% of children with language impairment, and Lugo-Neris et al. (2015) reported 95% sensitivity for the BESOS. However, specificity, or correct identification of typically developing children, for both screening measures falls below 80% (79% for the PLS-5 SST and 71% for the BESOS), indicating that these tools have a fairly high rate of overidentification of possible language impairment. Although specificity may not need to be as high for screening tools as for full diagnostic measures (Lugo-Neris et al., 2015), greater specificity of screening contributes to more efficient uses of resources as long as it does not occur at the expense of high sensitivity. For example, a screening tool with 95% sensitivity and 75% specificity would identify more children for referral for additional testing than a screening tool with 95% sensitivity and 85% specificity. From a clinical services point of view, high specificity also reduces the risk of unnecessary extensive testing and the potential for misidentification for individual children without language impairment.

A Bilingual DA Approach to Screening

Given the scarcity of screening tools to identify potential language impairment in DLLs, we have been working to create a tool tailored to address young bilingual children's linguistic diversity, their knowledge of two languages, and variation in sociocultural experiences. We developed dynamic language tasks that are administered in a brief, graduated prompting framework (Patterson et al., 2013). The tasks were developed taking a noncontrastive approach; that is, rather than developing tasks and items targeting each language separately, the tasks and items were developed in parallel, in English and Spanish, so that they would be appropriate in both languages. In view of the diversity of young bilingual children's exposure to and use of two languages, the examiner provides instructions, prompts, and models in the language that the child has heard at least half of the time or more, but the child is given positive feedback and credit for correct responses in either language. A noncontrastive approach to task and item selection, combined with bilingual scoring, may be an efficient way to sample children's overall language development without testing each language separately and then using combined scores or the "best score" across languages.

Because little is known about the use of graduated prompting for language screening purposes, our first study explored whether typically developing children's performance on brief tasks administered with graduated prompting would result in improvement within the task (Patterson et al., 2013). We found that performance of typically developing bilingual 4-year-olds improved on two language tasks when we compared their responses to the first two items versus the last two items on a set of brief, six-item tasks. We interpreted this as evidence that performance on these tasks (Novel Adjective Learning [NAL] and Similarity in Function [SF]) reflected learning or modifiability in response to support and that the tasks showed potential for screening. The focus of this study was to see if performance on tasks administered with graduated prompting would relate to children's later (year-end) performance on a language measure used by their preschool program. The year-end performance on developmental measures is an important metric for Head Start and other early intervention programs. If brief screening tools can identify children at risk for not achieving developmental expectations without additional supports, programs can provide comprehensive speech-language assessments and intervention services to the children who are most likely to benefit from those services. If performance on brief DA graduated prompting language tasks does relate to year-end language achievement, this approach would merit future studies focusing on sensitivity and specificity in identifying language impairment.

Selection of DA Tasks

Selection of tasks for DA presents several challenges at present. Tasks vary in a number of potentially important features that may affect their utility in the assessment of response to graduated prompting. These include the language domain under focus (e.g., vocabulary, narratives), task demands (e.g., learning new words or grammatical structures vs. prediction of future events), overall difficulty, and the types of prompts that are possible (e.g., focusing the child's attention on specific perceptual features vs. reminders of relevant information). In this multidimensional situation and given that relatively few tasks have been empirically explored in DA, differences in modifiability across tasks are not easily predictable or easily interpretable. Consequently, selection of tasks must represent a compromise between utilizing tasks that have been explored previously in DA along with new ones that reflect diversity in features. We selected three tasks: NAL, Prediction, and SF. The Prediction task was newly developed for this study, and the NAL and SF tasks were the two tasks from our previous work (Patterson et al., 2013) for which children showed evidence of learning across the six items within each task.

NAL is a receptive task. Novel word-learning tasks in which children are taught made-up words for unknown objects or actions have been widely used to examine language learning strategies in typically developing children's vocabulary learning as well as being applied in DA (Burton & Watkins, 2007). There is some evidence that novel word-learning tasks, in contrast with standard vocabulary tests, may diminish overidentification of language impairment in children from linguistic and cultural minority backgrounds, including children from diverse linguistic backgrounds in the United Kingdom (Camilleri & Law, 2014) and Spanish-speaking children in the United States (Kapantzoglou et al., 2012). In the present NAL task, rather than learning made-up words that might violate phonological constraints in one language or the other, children are taught real words in an existing language, Hawaiian, which has consonants and vowels that occur in English and Spanish phonological inventories and is not likely to be known by the vast majority of the participants. The children are taught adjectives such as "melemele" (yellow). The examiner labels pictures of yellow items among an array of yellow and pink items (*este es melemele y este es melemele…*—"this is melemele and this is melemele"), and the child is asked to point to the one that is melemele out of an array of new objects; if needed, this is followed by contrasting pictures and prompts that highlight the relevant feature (for a detailed example, see Patterson et al., 2013).

In the second DA task, Prediction, the child is asked to predict the results of an action, such as what will happen if more juice is poured into an already full glass (see an example in Appendix A). Although little, if any, work in DA has utilized this task, the importance of this skill is shown by the fact that tasks like this are included in language tests such as the Preschool Language Assessment Instrument–Second Edition (PLAI-2), which is designed to assess the ability of preschoolers to understand material at higher levels of abstractness—in this case, inferential skills that require going beyond what is shown or stated directly (Blank et al., 1978, 2003). In particular, this type of item is considered to be at the highest (fourth) level of abstraction, "Reasoning," on the PLAI-2. Preschool children with language impairments perform below the level of typically developing peers on inferential tasks such as this, and it has been suggested that this difficulty may contribute to the high risk of reading comprehension problems in the school years for these children (van Kleeck et al., 2006). Increasingly, prediction questions are included in language and literacy intervention programs. For example, as part of a family literacy intervention, Mesa and Restrepo (2019) trained Latino parents of Spanish-speaking preschool children to ask high-level questions, including prediction questions.

On our third task, SF, the child is shown a picture of two items with similar functions, such as a fork and a spoon, and asked how they are alike. As with the Prediction task, the SF task demands are based on the levels of abstraction in the PLAI-2. SF is classified at the third level of abstraction, "Reordering Perception" (Blank et al., 1978, 2003). The task also appears to align well with a reported tendency of Latino preschool children to focus on identifying object functions (Peña & Quinn, 1997). In summary, the three tasks were selected on the basis of relevance, previous research evidence, and diversity, rather than precise hypotheses about the relevance of the dimensions listed at the beginning of this section, which would have required restriction to systematic variation within a limited task domain.

Research Questions

The general aim of this study was to determine if performance on DA language tasks administered with graduated prompting would be related to year-end performance on language achievement testing among young DLLs from Spanish-speaking homes. Each of the DA language tasks included six individual items with three prompts per item that provided increasingly salient information to support the child's response. For the graduated prompting measure, we used scores from the last two items for each task. Because children have had minimal experience with a task on the initial items, we would not expect performance on the earlier items to reflect learning within the task. In contrast, scores from the last two items reflect children's performance after some experience with the task. Our (2013) finding of within-task improvement on the last two versus first two items recently was replicated on a phonological awareness task with fourth graders learning to read in English as a second (foreign) language in Taiwan (Lu & Hu, 2019). This pattern of improvement on different tasks, with children of different ages, and in different learning contexts indicates that performance on later items within tasks administered with graduated prompting reflects short-term learning and indirectly taps an aspect of modifiability.

In this study, the developmental achievement measure used by the participating Head Start programs, the Learning Accomplishment Profile–Third Edition (LAP3), was employed as a broad language measure to evaluate level of knowledge, as well as change over the year. The children's teachers administered this measure at the beginning of the year and at the end of the year. Although some stability would be expected, that is, a correlation of beginning and end of year scores, there will also be variability in progress. Some children may make rapid progress with preschool experience, while others may not. This is an especially relevant concern for children with lower LAP3 scores at the beginning of the year, as it is important to distinguish children who can be expected to gain from those who are more likely to continue to lag. To evaluate the extent to which performance on DA tasks can provide useful additional information beyond that provided by initial LAP3 language (LAP3-L) scores, we asked the following questions:

Question 1: Is children's performance on brief, DA language tasks related to their year-end LAP3-L scores?

Question 2: Does children's performance on DA tasks contribute to predicting year-end achievement (LAP3-L) beyond predictions based on initial LAP3-L scores?

Question 3: How much variance in year-end LAP3-L scores can be accounted for by the combination of initial LAP3-L scores and DA language task performance?

Method

This study (08-424/701337) was approved by the University of New Mexico Internal Review Board. Parents of the participants gave informed consent before they and their children participated in the study.

Participants

The participants were 20 four-year-olds (M_{age} = 54 months) from Spanish-speaking homes who were attending Head Start centers in a Southwestern metropolitan area of the United States. There were 10 boys and 10 girls in the study.

Spanish and English were used in the children's Head Start classes; all the Head Start classes had a Spanish-speaking teacher and/or classroom assistant. There were originally 30 children in the study, but we had initial and year-end LAP3 scores and at least 3 months between the graduated prompting language tasks and year-end LAP3 testing for only 20 of the children. We recruited participants from five preschool sites that the Head Start program director identified as having the most children from Spanish-speaking homes. The 20 children in this study were from nine different classrooms.

We recruited children from homes in which Spanish was spoken overall more than or about the same amount as English. For families that met this criterion, we asked parents to report on children's exposure to (input) and use of (output) each language with a 5-point scale previously used in large-scale investigations of young bilingual children's language development (Hammer et al., 2012). The five choices were as follows: all Spanish, more Spanish than English, about equal Spanish and English, more English than Spanish, and all English.

The children were from predominantly Spanish-speaking homes, as shown in Table 1. Five of the children's parents reported Spanish was the only language spoken in the home, and 14 reported more Spanish than English was spoken. Spanish and English were reportedly spoken about an equal amount in one child's home, but the mother reported she and the child spoke all Spanish; English and Spanish were used by older siblings and the other adult in the home. Thus, as would be expected among preschool DLLs, there was diversity in patterns of language input in the home, but all children shared the experience of hearing Spanish more than English in interactions with a primary caregiver, and except for one child from a home with about equal use of the two languages, all children heard more Spanish than English at home.

All of the children spoke at least some Spanish at home, and as shown in Table 1, the majority of the children (12) spoke all Spanish or more Spanish than English at home. However, four children reportedly spoke approximately equal amounts of Spanish and English, and four spoke more English than Spanish. Among these four children, two spoke more Spanish or all Spanish with the mother, one spoke equal amounts, and one spoke more English than Spanish with the mother (but the mother reported she spoke more Spanish to the child).

Teachers reported that all of the children heard both English and Spanish at school. Their reports using the

Table 1. Number of participants by levels of input and use of Spanish and English, based on parent and teacher report.

Context	All Spanish	More Spanish than English	About equal Spanish and English	More English than Spanish	All English
Home					
Child hears	5	14	1	0	0
Child speaks	3	9	4	4	0
School					
Child hears	0	3	6	11	0
Child speaks	1	6	6	4	3

5-point scale reflected this; none of the children was reported to hear "all Spanish" or "all English" at school. As shown in Table 1, three of the children were reported to hear more Spanish than English at school; two of these children were from monolingual Spanish homes, and one was from a home in which more Spanish than English was spoken. The children were from different classrooms, and other children in the same classrooms were estimated to hear different proportions of English and Spanish, so the reporting does appear to be based on teacher's perceptions of the individual children's experiences as we requested, rather than on the classroom as a whole. Teachers estimated approximately equal exposure to English and Spanish in the classroom for six of the children and more English than Spanish for the other 11 children. The children's use of Spanish and English in the classroom ranged from all English (three children) to all Spanish (one child), with the majority of the children using both languages to varying degrees.

We obtained demographic and educational background information from one parent (19 mothers and one father) in face-to-face or phone interviews. The reporting parents' education ranged from sixth grade to some college. The one father interviewed had completed high school. Among the mothers, seven had not completed high school, seven had completed high school (the median), and five had completed some college but did not have a degree.

Procedure

After an initial parent interview to obtain demographic and language experience information, we administered three dynamic language tasks. All testing was done by a bilingual graduate student in speech-language pathology. At the end of the school year, we obtained beginning and year-end language achievement testing results from the Head Start program.

Our intention was to conduct the dynamic language testing during the first 3 months of the school year. We were able to give half the children (10) the dynamic language tasks in the fall, as planned, but due to difficulties with logistics in recruitment and data collection, administration of the dynamic language tasks extended into winter for 10 of the children. Therefore, there was a range of 3–6 months between the time the dynamic language tasks were administered and when year-end testing was done.

Dynamic Language Assessment Tasks

Description

We administered three dynamic language assessment tasks. Two of the tasks, SF and NAL, had been used in a previous study, and bilingual children's within-task performance had improved significantly with support provided during task administration (Patterson et al., 2013). The SF task requires the child to identify similarities between two pictured items based on function (e.g., plastic spoon and fork). On the NAL task, we taught children Hawaiian adjectives (e.g., *poepoe*, which means "round"). The task is receptive as children are requested to point to a new object that is *poepoe* out of a set of three (one target and two foils). We included an additional task, Prediction, in which children were asked "What's going to happen?" when shown a picture of an impending situation (e.g., a person about to step in a puddle).

All tasks were developed in parallel in English and Spanish rather than starting with one language and then translating to the other. We did this to avoid developing items that had face validity in one language but were awkward in the other language.

Each of the tasks started with a demonstration item, followed by six items. During administration of each task, we provided feedback, modeling, and explanations. Correct responses were confirmed (e.g., "Yes, that one is *poepoe*"). Graduated prompts were provided following incorrect responses, and correct responses were modeled if the child did not respond correctly after two prompts. Prompts were designed to provide progressively more support, rather than being repetitions of the same prompt. An example from the Prediction task is provided in Appendix A. In general, the tasks are brief for children requiring minimal support to respond correctly, but they take somewhat longer for children who require more prompts.

Administration and Scoring

The tasks were administered in Spanish by a speech-language pathology graduate student. The student's native language was English, and her proficiency in Spanish, assessed as "advanced–low" on the Oral Proficiency Interview, was sufficient for accurate administration and scoring of the highly scripted tasks in this study, as indicated by reliability data (below). The task instructions, prompts, and feedback were given in Spanish because all of the children heard more Spanish than English at home and/or in interactions with their primary caregiver. However, correct responses in either language were credited because the focus was on the ability to understand and use language rather than on proficiency within each language. Children's responses were scored based on the number of prompts given. The maximum score on any item was "3" for a correct response on the initial presentation. A score of "2" was given for a correct response after the first prompt, a score of "1" was given for a correct response after two prompts, and a score of "0" was recorded if the child did not respond correctly after two successive prompts.

We used children's scores on the last two items on each task as a sample of optimal performance, that is, performance after the children had been provided with multiple supports/prompts, if needed, throughout the task. Performance on the last two items of tasks was partly analogous to year-end achievement testing, which took place after children's learning throughout the school year. Although posttesting with no teaching would have been more closely analogous to year-end achievement testing, the goal was to determine if it was possible to streamline assessment and forgo posttesting by using scores based on the last two items within the graduated prompting task. We did not use total task scores because performance on earlier items in the task most likely reflect prior experience. Children who do well on initial items presumably will continue to do well throughout the task, including on the last two items, particularly since they receive feedback confirming correct responses. On the other hand, among children who do not do well on initial items, performance on later items is a potential indicator of modifiability (or learning potential)—some children may improve in response to a brief session of teaching with graduated prompts and feedback, while others may not. We did not use change scores (difference between scores on the first two versus the last two items on each task) because children with high performance on the first two items have little to no opportunity to demonstrate improved performance on subsequent items. Because the possible scores on each item on the dynamic language assessment tasks range from 0 (*incorrect after two prompts*) to 3 (*correct without any prompts*), the range of possible scores for the last two items on a task was 0–6.

Reliability

Interscorer reliability for the dynamic language tasks was examined for 10 randomly selected children in the study. For the SF and Prediction tasks, children's responses were transcribed and listed in a random order for each item. Two transcripts were used; one was done by the original examiner, and the other was done by an undergraduate student who transcribed the responses from the video recordings. The transcribed responses then were scored independently by four bilingual graduate speech-language pathology students; two of the reliability scorers used the original transcript, and the other two reliability scorers used the second transcript. Each child response was then scored as correct or incorrect. We did this rather than have the scorers review the videos because the reliability scorers would know what the original score was based on whether or not the child was given prompts. Reliability scorers were provided the scoring manual, but they received no additional training. This was done so that the interscorer reliability checks would be done under conditions of minimal training, as might be expected if the tasks were used for screening purposes. Point-to-point agreement as "correct" or "incorrect" was 90% for the SF task for both scorers using the independent transcript and 93%–98% for scorers using the original transcript. Agreement on the Prediction task was 87%–95% with the independent transcript and 93%–97% for the original transcript. For NAL, which requires a pointing response, the reliability scores were based on one of the bilingual graduate students reviewing a video recording of the child's performance and scoring whether each of the child's responses was correct or incorrect; there was 98% agreement with the original scores.

LAP3

The Head Start programs the children attended used the LAP3, a criterion-referenced measure of developmental skills in seven domains. The LAP3 is individually administered by the child's teacher or assistant. We used the children's scores on the 69-item Language domain, which includes receptive and expressive items and print and letter recognition items. Examples include demonstrating understanding of four out of five prepositions (e.g., put the block *on/under/behind* the chair), telling the use of objects in response to questions (e.g., Why do we have keys?), and pointing to two out of three letters named by the teacher. The LAP3 manual (Hardin & Peisner-Feinberg, 2004) reports internal consistency for the language domain of .95 for children ages 48–53 months and .91 for children ages 54–59 months. Short-term (1–3 weeks) test–retest reliability of .96 is reported for the language domain, and interexaminer reliability is also high, that is, .93 for the language domain. The initial LAP3 assessments were done by the children's teachers in late October to mid-November, and the year-end assessments were completed about 6 months later, in April. LAP3 scores are the total number of items credited so the maximum score on the Language Scale (LAP3-L) is 69. Head Start staff credited the children with skills if they were demonstrated in Spanish or in English (or in both).

Results

Information on the length of time each task took and children's scores on the first two and last two items of each task are presented first, followed by results of the analyses related to the research questions and, finally, findings from follow-up analyses. The amount of time each task took varied with children's need for prompting. For NAL, the mean length of time was 4 min 47 s (range: 3:37–5:49), and length of time was inversely related to the child's total task score ($r = -.80$), indicating, as expected, that children with higher scores (i.e., those who needed less prompting) took less time. For Prediction, the mean length was 6 min 16 s (range: 4:00–12:15), and the correlation with the child's total score was –.87. For SF, the mean length was just over 6 min (range: 4:00–9:00), and the correlation with total task score was –.79.

Children's mean scores on the first two versus the last two items of each DA task were compared, and their performance on the initial LAP3-L was compared with their year-end LAP3-L performance. We used one-tailed tests for significance because we expected that children would, in each case, score higher on later items or, in the case of the LAP3-L, on year-end testing. We used Bonferroni corrections for the three DA tasks, but not for the LAP3-L, which was used in this study as an established criterion measure. As shown in Table 2, children's scores on the last two items were higher than scores on the first two items on all three dynamic language tasks, with the difference being statistically significant for the Prediction and SF tasks. Consistent with the analysis and results of the Patterson et al. (2013) study, we conducted an additional analysis for the NAL restricted to the 17 children who needed at least one prompt on the first item, and the resulting difference between the first two and last items was significant, $t(16) = 2.30$, one-tailed $p < .05$. As expected, the children's year-end LAP3 scores were significantly higher than their initial LAP3 scores at the beginning of the school year.

Our first research question was about relationships of children's performance on the DA tasks with year-end LAP3-L scores. Correlations of the year-end LAP3-L with the last two item scores for the three graduated prompting DA tasks (NAL, Prediction, and SF) were examined using one-tailed tests for significance because we predicted positive relationships. The results are shown in the first data column of Table 3. The findings were mixed; correlations were positive and significant for two out of the three DA tasks. Children's performance on the last two items of NAL tasks was not significantly correlated with LAP3-L year-end scores, but scores on the last two items of the Prediction and SF DA tasks were significantly and positively correlated with LAP3-L scores: for Prediction, $r = .49$, $p = .015$, and for SF, $r = .58$, $p < .005$. As shown in the second data column of Table 3, the findings were similar when controlling for the children's age.

Our second and third research questions concern incremental improvements in prediction of year-end LAP3-L scores. We utilized partial correlations to address these questions, as all the relevant variables (initial and year-end LAP3-L scores and DA task scores) require adjustment for age; multiple regression analyses would have adjusted only one variable at each step, as they utilize semipartial correlations. The second question was whether DA task performance was related to year-end LAP3-L scores when taking initial LAP3-L scores into account. As expected, children's year-end LAP3-L scores were significantly and positively related to their scores on the LAP3-L at the beginning of the school year ($r = .77$, $p < .001$), and this remained true when age was partialed out. As shown in the last column of Table 3, the relationship between children's performance on the SF task and LAP3-L scores remained significant when controlling for age and initial LAP3-L scores. This result is essentially a correlation between SF and change in LAP3-L over the year. The final question concerned the amount of variance in year-end LAP3-L scores that was accounted for by initial LAP3-L scores and DA task performance after controlling for age. Initial LAP3-L scores accounted for 50% ($.71^2$) of the variance in year-end LAP3-L scores (see Table 3, Column 2). SF, the DA task that did have a significant and unique relationship with year-end LAP3-L scores, accounted for an additional 23% ($.48^2$) of variance, as shown in Table 3, Column 3. Combined, the initial LAP3-L score and performance (last two items) on the SF task accounted for a total of 73% of the variance in year-end LAP3-L scores.

Follow-Up Analyses

Because SF is an expressive task, we conducted a follow-up analysis to determine if children's overall reported proportion of Spanish use was related to their SF task scores, even though responses in either language were credited. We combined reported amount of Spanish spoken by the child at home and at school, using the 5-point scale ranging from 1 = *all English* to 5 = *all Spanish*, so the range of

Table 2. Children's scores on dynamic assessment (DA) language tasks and the Learning Accomplishment Profile–Third Edition Language Scale (LAP3-L).

Measure	First two items *M* (*SD*)	Last two items *M* (*SD*)	Paired samples *t*
DA tasks (maximum score = 6)			
Novel Adjective Learning	3.58 (1.02)	4.21 (1.23)	$t(18) = 1.99$
Prediction	3.30 (2.25)	4.90 (2.02)	$t(19) = 3.65$**
Similarity in Function	2.25 (2.09)	3.90 (2.12)	$t(19) = 3.24$*
LAP3-L (maximum score = 69)	Initial 26.15 (16.19)	Year-end 40.25 (13.28)	$t(19) = 6.07$***

Note. *p* values with Bonferroni corrections are reported for the three DA tasks.
*$p < .01$. **$p < .005$. ***$p < .001$.

Table 3. Correlations of age, initial Learning Accomplishment Profile–Third Edition Language Scale (LAP3-L), and dynamic assessment (DA) language task scores with year-end LAP3-L scores.

Variable	Year-end LAP3-L	Year-end LAP3-L with age partialed out	Year-end LAP3-L with age and initial LAP3-L partialed out
Age	.64**		
LAP3–Language Initial	.77**	.71****	
DA–Novel Adjective Learning	−.32	−.25	−.21
DA–Prediction	.49*	.53*	.25
DA–Similarity in Function	.58***	.50*	.48*

*p < .05 (all are one-tailed). **p < .01. ***p < .005. ****p < .001.

possible ratings was 2 (*all English at home and at school*) to 10 (*all Spanish at home and at school*). We checked for potential relationships between Spanish spoken by the child and three different SF scores: total task score for SF (maximum possible = 18) and with scores on the first two and the last two items on the SF task. There was no correlation between reported proportion of Spanish spoken and the total SF task score (r = −.26, p = .27), nor with the first two SF items (r = −.36, p = .12) or the last two SF items (r = −.13, p = .58).

We also ran a one-way analysis of variance (repeated measures) to determine if there were differences in difficulty among the tasks. To do this, we compared children's responses to the first two items of NAL, Prediction, and SF to determine if there were significant differences early in the tasks. There were significant differences (F = 8.39, p < .01) among the task scores, and post hoc contrasts indicated scores on the first two items of the SF task were significantly lower than scores on the first two items for the NAL and Prediction tasks (p < .01 for SF vs. NAL and for SF vs. Prediction).

Discussion

The results of this study indicate that a graduated prompting approach to DA holds promise for potential screening purposes; however, the variation in findings across DA tasks indicates that using this approach to DA requires careful consideration of the content and demands of individual tasks. Therefore, task-specific findings and considerations are discussed first, followed by discussion of implications for using appropriate graduated prompting tasks for screening for language impairment in young DLLs.

Children's Performance on the Three DA Language Tasks

Two of the three DA language tasks we used in this study were related to subsequent performance on a static measure of children's language skills. The two tasks required verbally identifying similarity in function between two objects (SF) and predicting impending action given a visual scenario (Prediction). For these tasks, children appeared to "learn the game" sufficiently to need fewer prompts on the last two items compared to the initial two items on the six-item tasks, and their scores on the last two items related significantly to year-end language measures. Furthermore, the significant relationship of performance on the SF graduated prompting task with year-end language testing (LAP3-L) was significant, even when controlling for initial language scores on the LAP3-L at the beginning of the year. The relationship of performance on the Prediction task, on the other hand, was no longer significant when beginning-of-the-year testing was taken into account.

In contrast with the SF and Prediction tasks, children's performance on the NAL task did not improve significantly from the first two items to the last two items for the group as a whole. Similar to findings in the Patterson et al. (2013) study, when looking specifically at children who required at least one prompt on the first item on the NAL task, scores on the last two items were significantly higher than on the first two items. Although there was improvement from early to late items on the NAL task among children who needed at least some prompting initially, performance on the last two NAL items was not positively related to performance on the year-end static language measure.

Among the three tasks used in this study, children's performance on the SF task showed the strongest potential for use in screening. This may be attributed to the nature of the task demands and the level of difficulty of the SF task.

Task Demands

On the SF task, the focus consistently was on identifying the similarity in function. Similarly, the Prediction task consistently required children to predict the consequence of an impending action (e.g., stepping into a puddle). The NAL task required that children identify a feature and link it to a novel adjective based on presentation of a set of exemplars contrasting on the relevant feature (e.g., green vs. white items). In contrast with the consistent focus within the SF and Prediction tasks, the feature that children needed to focus on changed across items on the NAL task: For two of the NAL items, the quality identified by the novel adjective is a color; for three others, it is shape; and for another, it is whole versus broken.

In addition, NAL and other novel word-learning tasks require that children learn a new word for each item. Given that each item requires specific learning, it is possible that overall performance on novel word-learning tasks is best represented by a total score, reflecting children's responses to all the words taught, in contrast with the scoring method used in this study, which focused on the last two items in the task.

It may also be relevant that our NAL task differed from the format of most other word-learning tasks. Rather than presenting novel objects and requiring the child to learn a label for the object or some perceptual or functional quality, our task required the child to learn in effect a new label for an existing concept. Although this might be a desirable feature for purposes of assessing multilingual learning abilities, it does differ from novel word-learning tasks. Given the current findings, it appears that other novel word-learning paradigms hold more promise for identifying possible language impairment. Furthermore, including expressive tasks in screening protocols may be more effective than using strictly receptive tasks such as the NAL task. In a study that included receptive and expressive novel vocabulary learning by preschoolers, only an expressive measure was predictive of future vocabulary learning (Camilleri & Law, 2014). Scoring method may also be relevant; a combination of receptive and expressive word-learning scores and examiner rating of children's learning strategies resulted in moderate to good identification of predominantly Spanish-speaking preschool children with and without language impairment (Kapantzoglou et al., 2012). Both studies also employed scores based on children's performance on all of the novel words taught rather than scores being based on performance on the last two items, as was done in this study.

Task Difficulty

Although the SF and Prediction tasks were similar in requiring a consistent within-task focus, the Prediction task may have been too easy for the group as a whole. The mean score on the initial two items of the SF task was significantly lower than the mean score on the first two items of both NAL and Prediction. Looking at individual scores for the first two items of the Prediction task, six of the 20 children got the maximum possible score (6/6) on the first two items; only one child scored 6/6 on the initial items on the SF task. Thus, for the Prediction task, 30% of the children were able to do the task independently and accurately on the initial items, so the task could not assess responses to support or modifiability for almost one third of the sample. Given that a sizable portion of the children did not have the opportunity to demonstrate learning due to the task being insufficiently challenging for them, it is not surprising that Prediction failed to contribute to predicting future language achievement on the year-end static language achievement measure (LAP3-L) for the group as a whole. A more challenging set of Prediction items might have greater predictive validity.

Potential for a Graduated Prompting Approach to DA for Screening

Children's performance on the DA tasks showed evidence of short-term learning, based on their within-task improvement from the first two to the last two items when given prompts and feedback. For children who did not achieve the maximum score on the initial task items, scores on the last two items reflected, in part, decreases in need for support after experience with a demonstration item plus feedback and prompting on four prior items within the task. These scores represent responsiveness to brief but concentrated support, perhaps parallel to the longer range outcomes of children's responses to several months of broader support provided in the classroom. In fact, children's performance on the SF and Prediction DA tasks was related to the program's year-end language achievement testing. The potential of the SF task was particularly strong based on its relationship with year-end LAP3-L scores, accounting for almost a quarter of the variance in year-end LAP3-L scores beyond the variance in LAP3-L scores accounted for by age and initial LAP3-L scores.

Taken together, the results of this study and of a prior study using graduated prompting tasks (Patterson et al., 2013) contribute some evidence of validity of one type of DA, partially addressing the call for different lines of evidence to test whether DA procedures are tapping modifiability/learning potential (Poehner, 2011). Specifically, children's performance on graduated prompting DA tasks demonstrates that learning can take place within the constraints of a brief session with scripted prompts (Patterson et al., 2013), and this study shows that performance on one of these DA tasks, averaging 6 min in length, is a significant predictor for performance on a language achievement measure (the LAP3) administered 3–6 months later.

It is important to note that assessing children's responses to more extensive support is important in full diagnostic assessment. For example, Petersen et al. (2017) achieved 100% sensitivity and 88% specificity in one 25-min session when using a DA narrative assessment with kindergarten through third-grade children; both sensitivity and specificity were 100% when performance across two 25-min sessions was examined. The brief, scripted graduated prompting tasks we used in a single session in this study were specifically designed for potential screening applications.

In spite of the constraints of a graduated prompting approach, there was a significant relationship with children's later performance on language achievement testing, particularly for the SF task. This may be due, in part, to the incorporation of best assessment practices for addressing linguistic and sociocultural diversity. Practices focusing on linguistic diversity included developing the DA tasks with both languages in mind, administering them in the appropriate language, and using bilingual scoring. Children's language achievement by the children's Head Start teachers also was assessed using bilingual scoring. The use of bilingual scoring allowed credit for correct responses in either language, an important feature since young bilingual children's scores on language tasks are related to a wide variety of variables, including cumulative and current exposure to each language and, for expressive tasks, the degree to which the child uses each language (Bohman et al., 2010; Hammer et al., 2012; Ribot et al., 2018).

Children's diverse learning experiences associated with differences in social, economic, and cultural backgrounds were addressed by using a DA graduated prompting framework. In addition to these practices, however, testing the individual tasks revealed that the task demands must be carefully considered when designing DA tasks for use with graduated prompting. Of the three tasks tested in this study (NAL, Prediction, and SF), the SF and Prediction tasks showed within-task improvement and were related to year-end language achievement testing; in addition, the SF task accounted for a substantial amount of variance in later language achievement testing while controlling for children's beginning-of-the-year language achievement scores.

Although the sample size for this study was small and especially so for children with low initial LAP3-L scores, the relationship seen here between the SF DA task and year-end LAP3-L scores is encouraging. An essential next step will be to explore what, if any, cut-point for the last two item scores on the SF task would yield acceptable accuracy of identification of children at risk for low achievement and/or language impairment. This is particularly important in view of the interindividual variation among children, as shown in Appendix B. It is possible that adding brief DA tasks such as SF for children achieving low scores on static language tests at the beginning of the school year would help preschool programs determine which of those children should be referred for further testing to identify possible language impairment and need for specialized speech-language intervention services, but this is an open question requiring further research.

Limitations and Future Directions

Limitations of this study include the tool and procedures used by the children's preschool program to measure language skills, the time span between dynamic tasks and year-end measures, and the sample size. First, although responses in either language were accepted on the year-end language measure, the LAP3-L scale is an English-based measure. This means that some important skills and structures in Spanish likely were not included in the measure used by the program to document skills and progress in language development during the school year. Furthermore, although the Head Start administration instructed teachers to credit items on the LAP3-L if the child responded correctly in either language, there most likely was variation in how this directive was carried out by individual teachers. Although this potential source of variability is a limitation, it is reflective of real-world practices rather than controlled testing conditions, and it is encouraging that, nonetheless, children's performance on the SF DA task was significantly related to their future performance on the program's independent year-end static testing.

Ideally, screening would occur at the beginning of the year; however, due to a number of logistic challenges, administration of potential DA language screening tasks stretched into the second half of the school year in this study. We excluded children with less than 3 months between DA task administration and year-end LAP3-L testing, so the mean time between DA and year-end LAP3-L was 4.2 months; the interval between testing was 3–4 months for half the children and 5–6 months for the other 10 children.

The size and nature of the participant sample in this study is a significant area of limitation. The small sample provided very limited statistical power for identifying or ruling out multivariate relationships, including performance on the Prediction task with year-end LAP3-L performance when controlling for initial LAP3-L scores and the children's exact age. In addition, the children were all from five Head Start preschools in a single metropolitan area. The Head Start preschool programs were committed to a dual language philosophy and had at least one Spanish-speaking teacher or assistant in each classroom. The sample size and shared experience of attending Head Start preschools within a dual language framework constrain the potential for generalization of the study findings to other young DLLs.

A limitation that is inherent to all DA research at present is the lack of knowledge about the reliability and validity of proposed DA tasks. What is especially needed is systematic exploration of tasks, which can illuminate the role of specific features of DA tasks such as language domain, task demands, and nature of the prompts in determining the modifiability of tasks and the validity of modifiability in these miniature learning situations as predictors of benefit from more intensive and sustained educational intervention. It is intriguing, though far from definitive, that the most successful task utilized here, SF, was characterized by consistency of task demand, moderate initial difficulty, connection to real-world knowledge, and prompts that provided verbal and visual cues highlighting the shared functions of objects.

In spite of the limitations summarized above, the young DLLs in this study did show gains on brief, graduated prompting DA tasks, and their performance on the later items on one of the tasks (SF) was related to subsequent language achievement even when controlling for scores on initial static language achievement testing. Given these findings, a graduated prompting approach to DA for screening purposes appears to be promising when carefully developed and tested tasks are used. Given that the mean time for administering the most successful task in this study, SF, was 6 min and the length of time was related to the need for prompting, it may be that brief DA tasks could be used as a supplemental tool for screening since the task would take little time for children who needed minimal support and somewhat more time for children requiring multiple prompts.

Studies with larger and geographically more diverse samples are needed to determine the generalizability of the findings and to allow more in-depth task contrasts and analyses of task performance. For example, systematically testing the predictive capacity of specific tasks for children with low performance was not done in this study due to the small sample size. In addition, a priority for future research is developing additional tasks or modifying tasks to fit the practical and conceptual parameters involved in applying a graduated prompting framework for screening purposes. For tasks such as SF that have met the criteria of being related to future performance beyond the relationship of initial and year-end static testing, a key step is to evaluate whether performance on carefully developed and tested DA language tasks administered with graduated prompting can accurately identify likelihood of language impairment. Although screening using DA language tasks could be applied as a universal screening tool, perhaps brief DA tasks could be used as a follow-up for children who score low on initial static testing in order to determine if a referral for more comprehensive language assessment is warranted. Ultimately, development of tools using appropriate tasks for a graduated prompting approach to DA may contribute to solving the problems of under- and over-identification of language impairment in young DLLs and thereby contribute to providing appropriate intervention services for preschool children with and without language impairment.

Acknowledgments

Many thanks to the Youth Development Inc Head Start program directors and teachers and to the families who participated in this study. Many Speech and Hearing Sciences students contributed to this study: Kathryn Mancewicz was responsible for recruitment, data collection, and data entry. Her persistence and dedication were essential to the completion of the project. Mireya Hernandez and Nykki Montaño worked on designing the Prediction task. Michael Campbell arranged the reliability studies and helped with formatting the references and appendixes. Others helped with reliability transcription and scoring (Susan Carrasco, Cruz Martinez, and Nydia Ramírez), and Rikki Girón helped with recruiting, parent interviews, and reliability scoring.

References

Anaya, J. B., Peña, E. D., & Bedore, L. M. (2018). Conceptual scoring and classification accuracy of vocabulary testing in bilingual children. *Language, Speech, and Hearing Services in Schools, 49*(1), 85–97. https://doi.org/10.1044/2017_LSHSS-16-0081

Blank, M., Rose, S. A., & Berlin, L. J. (1978). *Preschool language assessment instrument: The language of learning in practice*. The Psychological Corporation.

Blank, M., Rose, S. A., & Berlin, L. J. (2003). *PLAI-2: Preschool Language Assessment Instrument–Second Edition*. Pro-Ed.

Bohman, T., Bedore, L., Peña, E., Mendez-Perez, A., & Gillam, R. (2010). What you hear and what you say: Language performance in Spanish English bilinguals. *International Journal of Bilingual Education and Bilingualism, 13*(3), 325–344. https://doi.org/10.1080/13670050903342019

Bridges, M. S., & Catts, H. W. (2011). The use of a dynamic screening of phonological awareness to predict risk for reading disabilities in kindergarten children. *Journal of Learning Disabilities, 44*(4), 330–338. https://doi.org/10.1177/0022219411407863

Burton, V. J., & Watkins, R. (2007). Measuring word learning: Dynamic versus static assessment of kindergarten vocabulary. *Journal of Communication Disorders, 40*(5), 335–356. https://doi.org/10.1016/j.jcomdis.2006.06.015

Camilleri, B., Hasson, N., & Dodd, B. (2014). Dynamic assessment of bilingual children's language at the point of referral. *Educational and Child Psychology, 31*(2), 57–72.

Camilleri, B., & Law, J. (2014). Dynamic assessment of word learning skills of pre-school children with primary language impairment. *International Journal of Speech-Language Pathology, 16*(5), 507–516. https://doi.org/10.3109/17549507.2013.847497

Gutierrez-Clellen, V. F., & Kreiter, J. (2003). Understanding child bilingual acquisition using parent and teacher reports. *Applied Psycholinguistics, 24*(2), 267–288. https://doi.org/10.1017/S0142716403000158

Hammer, C. S., Komaroff, E., Rodríguez, B. L., Lopez, L. M., Scarpino, S. E., & Goldstein, B. (2012). Predicting Spanish–English bilingual children's language abilities. *Journal of Speech, Language, and Hearing Research, 55*(5), 1251–1264. https://doi.org/10.1044/1092-4388(2012/11-0016)

Hammer, C. S., & Rodríguez, B. L. (2012). Bilingual language acquisition and the child socialization process. In B. Goldstein (Ed.), *Bilingual language development and disorders in Spanish–English speakers* (2nd ed., pp. 31–46). Brookes.

Hardin, B., & Peisner-Feinberg, E. (2004). *The Learning Accomplishment Profile–Third Edition (LAP-3) examiner's manual & technical report*. Kaplan Early Learning. https://www.kaplanco.com/media/LAP3_Manual.pdf

Henderson, D. E., Restrepo, M. A., & Aiken, L. S. (2018). Dynamic assessment of narratives among Navajo preschoolers. *Journal of Speech, Language, and Hearing Research, 61*(10), 2547–2560. https://doi.org/10.1044/2018_JSLHR-L-17-0313

Hibel, J., & Jasper, A. D. (2012). Delayed special education placement for learning disabilities among children of immigrants. *Social Forces, 91*(2), 503–530. https://doi.org/10.1093/sf/sos092

Kapantzoglou, M., Restrepo, M. A., & Thompson, M. S. (2012). Dynamic assessment of word learning skills: Identifying language impairment in bilingual children. *Language, Speech, and Hearing Services in Schools, 43*(1), 81–96. https://doi.org/10.1044/0161-1461(2011/10-0095)

Lu, Y.-Y., & Hu, C.-F. (2019). Dynamic assessment of phonological awareness in young foreign language learners: Predictability and modifiability. *Reading and Writing, 32*(1–2), 891–908. https://doi.org/10.1007/s11145-018-9894-x

Lugo-Neris, M. J., Peña, E. D., Bedore, L. M., & Gillam, R. B. (2015). Utility of a language screening measure for predicting risk for language impairment in bilinguals. *American Journal of Speech-Language Pathology, 24*(3), 426–437. https://doi.org/10.1044/2015_AJSLP-14-0061

Mesa, C., & Restrepo, M. A. (2019). Effects of a family literacy program for Latino parents: Evidence from a single-subject design. *Language, Speech, and Hearing Services in Schools, 50*(3), 356–372. https://doi.org/10.1044/2018_LSHSS-18-0035

Paradis, J., Genesee, F., & Crago, M. B. (2011). *Dual language development and disorders* (2nd ed.). Brookes.

Park, M., O'Toole, A., & Katsiaficas, C. (2017). *Dual language learners: A national demographic and policy profile*. Migration Policy Institute.

Patterson, J. L., Rodríguez, B. L., & Dale, P. S. (2013). Response to dynamic language tasks among typically developing Latino preschool children with bilingual experience. *American Journal of Speech-Language Pathology, 22*(1), 103–122. https://doi.org/10.1044/1058-0360(2012/11-0129)

Pearson, B. Z. (1998). Assessing lexical development in bilingual babies and toddlers. *International Journal of Bilingualism, 2*(3), 347–372. https://doi.org/10.1177/136700699800200305

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

Peña, E. D., Gillam, R. B., & Bedore, L. M. (2014). Dynamic assessment of narrative ability in English accurately identifies language impairment in English language learners. *Journal of Speech, Language, and Hearing Research, 57*(6), 2208–2220. https://doi.org/10.1044/2014_JSLHR-L-13-0151

Peña, E. D., Iglesias, A., & Lidz, C. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10*(2), 138–154. https://doi.org/10.1044/1058-0360(2001/014)

Peña, E. D., & Quinn, R. (1997). Task familiarity: Effects on the test performance of Puerto Rican and African American children. *American Language, Speech, and Hearing Services in Schools, 28*(4), 323–332. https://doi.org/10.1044/0161-1461.2804.323

Petersen, D. B., Changthongthip, H., Ukrainetz, T. A., Spencer, T. D., & Steeve, R. (2017). Dynamic assessment of narratives: Efficient, accurate identification of language impairment in bilingual students. *Journal of Speech, Language, and Hearing Research, 60*(4), 983–998. https://doi.org/10.1044/2016_JSLHR-L-15-0426

Poehner, M. E. (2011). Validity and interaction in the ZPD: Interpreting learner development through L2 dynamic assessment. *International Journal of Applied Linguistics, 21*(2), 244–263. https://doi.org/10.1111/j.1473-4192.2010.00277.x

Poehner, M. E., & Lantolf, J. P. (2013). Bringing the ZPD into the equation: Capturing L2 development during Computerized Dynamic Assessment (C-DA). *Language Teaching Research, 17*(3), 323–342. https://doi.org/10.1177/1362168813482935

Resing, W. C. M. (1997). Learning potential assessment: The alternative for measuring intelligence? *Educational and Child Psychology, 14,* 68–82.

Ribot, K. M., Hoff, E., & Burridge, A. (2018). Language use contributes to expressive language growth: Evidence from bilingual children. *Child Development, 89*(3), 929–940. https://doi.org/10.1111/cdev.12770

Rogoff, B. (2003). *The cultural nature of human development*. Oxford University Press.

Samson, J. F., & Lesaux, N. K. (2009). Language-minority learners in special education: Rates and predictors of identification for services. *Journal of Learning Disabilities, 32*(2), 148–162. https://doi.org/10.1177/0022219408326221

Scarpino, S. E., Hammer, C. S., Goldstein, B., Rodríguez, B. L., & Lopez, L. M. (2018). Effects of home language, oral language skills, and cross-linguistic phonological abilities on whole-word proximity in Spanish-English–speaking children. *American Journal of Speech-Language Pathology, 28*(1), 174–187. https://doi.org/10.1044/2018_AJSLP-18-0050

Seymour, H. N., Roeper, T. W., & de Villiers, J. (2005). *Diagnostic Evaluation of Language Variation–Norm Referenced: Examiner's manual*. PsychCorp.

Tzuriel, D. (2000). Dynamic assessment of young children: Educational and intervention perspectives. *Educational Psychology Review, 12*(4), 385–435. https://doi.org/10.1023/A:1009032414088

U.S. Census Bureau. (2015). *American community survey reports: Language use in the United States*.

van Kleeck, A., Vander Woude, J., & Hammett, L. (2006). Fostering literal and inferential language skills in Head Start preschoolers with language impairment using scripted book-sharing discussions. *American Journal of Speech-Language Pathology, 15*(1), 85–95. https://doi.org/10.1044/1058-0360(2006/009)

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

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2015). *Preschool Language Scales–Fifth Edition Spanish Screening Test*. Pearson.

Appendix A (p. 1 of 3)

Prediction Task Example

ENGLISH VERSION

TASK INTRODUCTION AND DEMONSTRATION ITEM: (Tomato cans)

- Examiner*:* **"We're going to look at some photos and think about what might happen next. Look at this girl."**
- *(Child is shown a color photograph of the cans of tomatoes stacked in a pyramid and a girl about to pull out the bottom of the stack.)*
- Examiner: **"What would happen to this can** (*points to can on top)* **if she removes this can** (*points to can on bottom)?*"
- Child Response
  - If correct (e.g., it would fall, it'll fall), examiner says: **"That's right. If she pulls the can out of the bottom** (*points to bottom can*), **then this can** (*points to top can)* **would fall. That's what would happen."**
  - If incorrect: **"If she pulls out the bottom can** (points to bottom can), **then this can** (points to top can) **would fall. That's what would happen."**

Appendix A (p. 2 of 3)

Prediction Task Example

FIRST TASK ITEM: (<u>puddle</u>)

- Examiner: **"Let's try another one."**
- *Examiner shows child a color photograph of a man about to step in a puddle of water.*
- Examiner*:* **"Look at this man. What will happen to his shoes if he steps in the puddle?"**
- Child response
  - If correct: **"That's right! If the man steps in the puddle, his shoes will get wet. That's what would happen."** ***Proceed to next item*.**
  - If incorrect, examiner says **"Hmmm…Let's look again"** and proceeds to Prompt 1.

Prompt 1 for puddle:

- Examiner **shows child the same photograph** and says, **"Look at the puddle** (points to puddle). **What will happen to his shoes if he steps in the puddle?"**
  - If response is correct: **"That's right! If the man steps in the puddle, his shoes will get wet. That's what would happen."** ***Proceed to next item*.**
  - If response is incorrect, examiner says **"Let's look again more closely"** and proceeds to Prompt 2.

Prompt 2 for puddle:

- Examiner *shows child a close-up color photograph* of the same man's shoe/foot about to step in a puddle of water.
- Examiner says *and points* **"Look at the puddle…"**
- Examiner *returns to the original photograph of the man about to step in a puddle* and says: **"What will happen to his shoes if he steps in the puddle?"**
  - If response is correct: **"That's right! If the man steps in the puddle, his shoes will get wet. That's what would happen."**
  - If response is incorrect, examiner models the correct response **"If the man steps in the puddle, his shoes will get wet. That's what would happen."**

<u>SPANISH VERSION</u>

TASK INTRODUCTION AND DEMONSTRATION ITEM: (<u>Tomato cans</u>)

- Examiner: **"Vamos a mirar a unas fotos y pensar en que va a pasar después/enseguida."**
- Child is shown a color photograph of the cans of tomatoes stacked in a pyramid and a girl about to pull out the bottom of the stack.
- Examiner: **Mira a esta muchacha…¿Qué va a pasar con/a esta lata** *(points to top can)* **si quite/saque esta lata?** *(points to bottom can)."*
- Child Response
  - If correct (e.g., se va a caer; se cae). Examiner says: **"¡Eso es! Si ella quitará/sacará la lata de abajo** *(points to can),* **entonces esta lata** *(points to top can)* **caería. ¡Esto es lo que pasaría!"**
  - If child's response is incorrect: **"Si ella quitará/sacará la lata de abajo** *(point to bottom can),* **entonces esta lata** (*point to top can)* **caería. ¡Esto es lo que pasaría!"**

FIRST TASK ITEM: (<u>puddle</u>)

- Examiner: **'Mira a este hombre..."**
- Examiner shows child a color photograph of a man about to step in a puddle of water.
- Examiner: **"¿Qué les va a pasar a los zapatos si pise en el charco?"**
- Child response:
  - If correct: **"¡Eso es! Si el hombre pise en el charco, sus zapatos van a mojarse. ¡Eso es lo que pasaría!"**
  - If incorrect, examiner says **"Hmm… vamos a mirarlo otra vez"** and go to Prompt 1.

Appendix A (p. 3 of 3)

Prediction Task Example

Prompt 1 for puddle:

- Examiner shows child the same photograph and says, "**Mira al charco** *(points to puddle)*. **Qué les va a pasar a los zapatos si pise en el charco?"**
    - If response is correct: **"¡Eso es! Si el hombre pise en el charco, sus zapatos van a mojarse. ¡Eso es lo que pasaría!"**
    - If response is incorrect, examiner says **"Miramos otra vez con más cuidado"** and proceed to Prompt 2

Prompt 2 for puddle:

- Examiner shows child a close-up color photograph of the same man's shoe/foot about to step in a puddle of water.
- Examiner says and points "**Mira al charco…"**
- Examiner returns to the original photograph of the man about to step in a puddle and says: **"¿Qué les va a pasar a los zapatos si pise en el charco?"**
    - If response is correct: **"¡Eso es! Si el hombre pise en el charco, sus zapatos van a mojarse. ¡Eso es lo que pasaría!"**
    - If response is incorrect, examiner models the correct response **"Si el hombre pise en el charco, sus zapatos van a mojarse. Eso es lo que pasaría."**

Appendix B

Individual Children's Scores on Initial and Year-End Learning Accomplishment Profile–Third Edition Language Scale (LAP3-L) and First Two and Last Two Items on the Similarity in Function (SF) Dynamic Assessment Task

Participant	LAP3-L initial	LAP3-L year-end	SF first two items	SF last two items
01	7	50	1	6
02	44	53	0	4
03	22	41	3	3
04	54	59	6	6
06	42	55	0	5
07	17	41	3	4
09	54	58	5	6
10	12	32	5	6
11	46	46	2	4
12	32	42	3	6
15	27	40	1	2
16	31	57	4	6
17	41	42	3	3
18	9	11	0	0
19	12	36	0	0
22	13	23	4	3
23	11	31	0	6
24	11	24	0	5
25	31	41	5	3
26	7	23	0	0

Note. Gaps in participant numbers are due to excluding participants for whom we did not have year-end LAP3-L scores or for whom there was less than 3 months between SF task administration and year-end LAP3 testing.