Predicting Reading Ability for Bilingual Latino Children Using Dynamic Assessment

Douglas B. Petersen, PhD¹ and Ronald B. Gillam, PhD²

Abstract
This study investigated the predictive validity of a dynamic assessment designed to evaluate later risk for reading difficulty in bilingual Latino children at risk for language impairment. During kindergarten, 63 bilingual Latino children completed a dynamic assessment nonsense-word recoding task that yielded pretest to posttest gain scores, residuum gain scores, and modifiability scores. At the end of first grade, the same participants completed criterion reading measures of word identification, decoding, and reading fluency. The dynamic assessment yielded high classification accuracy, with sensitivity and specificity at or above 80% for all three criterion reading measures, including 100% sensitivity for two out of the three first-grade measures. The dynamic assessment used in this study has promise as a means for predicting first-grade word-level reading ability in Latino, bilingual children.

Keywords
assessment, reading, early identification/intervention, cultural and linguistic diversity

Latino children comprise a large percentage of the millions of school-age students in the United States who are classified as having reading difficulty. According to the National Assessment of Educational Progress (NAEP; 2011), 82% of fourth-grade Latino children read below a proficient level, with nearly 50% reading below a basic level. This high percentage of Latino children who are struggling readers is greatly concerning. Reading is a fundamental skill required for academic achievement and is essential for success in today’s society (Snow, Burns, & Griffin, 1998).

Reading difficulty, once manifest, is pervasive. Nearly 75% of first-grade students who are at risk for reading problems will continue to have difficulty reading into adulthood (Lyon, 2004; Scarborough, 1998). Researchers have demonstrated that reading problems can be corrected at an early age when evidence-based practices are applied (Brown & Felton, 1990; Elbro & Scarborough, 2004; Foorman, Francis, Fletcher, Schatschneider, & Mehta, 1998; Justice & Pullen, 2003; Morris, Tyner, & Perney, 2000; National Reading Panel [NRP], 2000; Torgesen et al., 2001; Vaughn et al., 2006; Vellutino, Scanlon, & Tanzman, 1998). The early, accurate identification of children at risk for reading difficulty should be one of the first steps in preventing reading problems from manifesting.

Any assessment that is designed to classify children into ability groups should have adequate sensitivity and specificity. In relation to the identification of future reading difficulty, sensitivity refers to the ability of a measure to correctly identify children at risk for reading difficulty (true positives). Specificity refers to the correct identification of children who are not at risk for reading difficulty (true negatives). Plante and Vance (1994) suggest that tests meet a minimum criterion of .80 sensitivity and specificity.

The majority of research pertaining to the predictive evidence of validity of early reading measures has revealed that a balance between good sensitivity and good specificity is difficult to obtain (Catts, Petscher, Schatschneider, Bridges, & Mendoza, 2009; Elbro & Scarborough, 2004; Hintze, Ryan, & Stoner, 2003; Johnson, Jenkins, Petscher, & Catts, 2009; Scarborough, 1998; Snow et al., 1998; Speece, Mills, Ritchey, & Hillman, 2003). Most research has focused on the administration of static measures to young, English-speaking, monolingual children (Klingner & Edwards, 2006), and such studies have often overpredicted risk status, reporting false positives ranging from 47% to 67% (O’Connor & Jenkins, 1999; Uhry, 1993) and high rates of false negatives (under-identification) ranging from 21% to 69% (Coleman & Dover, 1993; Mantzicopoulos & Morrison, 1994; Torgesen, Burgess, Wagner, & Rashotte, 1996).

¹University of Wyoming, Laramie, WY, USA
²Utah State University, Logan, UT, USA

Corresponding Author:
Douglas B. Petersen, University of Wyoming, Department 3311, 1000 E. University Ave., Laramie, WY 82071-2000, USA.
Email: dpeter39@uwyo.edu
Research on Latino English Language Learners

For Latino English Language Learners (ELLs), measures administered in English and in Spanish have accounted for unique variance in English reading component skills (Lindsey, Manis, & Bailey, 2003; Manis, Lindsey, & Bailey, 2004; Páez & Rinaldi, 2006; Swanson, Sáez, Gerber, & Lefstead, 2004). These research findings indicate that Spanish-related pre-literacy measures and Spanish language ability are predictive of reading in English for Latino ELLs to some degree, yet sensitivity and specificity have not been high for any combination of English and/or Spanish predictor measures. For example, Lindsey et al. (2003) assessed 249 Spanish-speaking ELLs at kindergarten using Spanish-only static measures. They tested the children again at the end of first grade measuring decoding in English and word identification and reading comprehension in English and Spanish, with results yielding sensitivity ranging from .63 to .77 and specificity ranging from .60 to .78. Likewise, Vanderwood, Linklater, and Healy (2008) administered a static nonsense word fluency measure at first grade to predict the third-grade reading ability of 134 ELLs (90% Latino). The researchers reported specificity values above .70 but sensitivity values at or below .50. These ranges of accuracy are generally considered unacceptable (Carran & Scott, 1992; Gredler, 2000; Kingslake, 1983) and indicate that a relatively large percentage of ELL children were over- or underidentified for being at risk for reading difficulty.

There are several possible reasons why the predictive validity of static, single-time measures of reading is poor. When assessing ELL children, oral language proficiency, prior instruction, socioeconomic status (SES), and cultural differences could significantly affect the outcomes of the assessment. Because of these factors, static assessments have a high potential for cultural and linguistic bias and measurement error (Bracken, 1998; Densky, Mittenberg, Quintar, Katell, & Golden, 1998; McCauley & Swisher, 1984a, 1984b; Plante & Vance, 1994; Rodekohr & Haynes, 2001; Scheffner-Hammer, Pennock-Roman, Rzasa, & Tomblin, 2002; Spaulding, Plante, & Farinella, 2006; Valencia & Rankin, 1985; Valencia & Suzuki, 2001). Static measures are currently not meeting the technical requirements needed for the valid early identification of reading difficulty, and with an ever increasing population of culturally and linguistically diverse children, assessments that yield less measurement and contextual bias are needed.

Dynamic Assessment

Vellutino et al. (1998) argued against the exclusive use of static assessment measures for assessing reading difficulty in culturally and linguistically diverse individuals. They suggested that most reading difficulties are caused by inadequate literacy experiences and inadequate instruction (for related findings, see August & Shanahan, 2006; Clay, 1987; Kim & Suen 2003; Stanovich, 1986). Vellutino et al. remarked that perhaps the best way to determine whether a child has a reading problem because of experiential or instructional deficits (or because of limited English ability) as opposed to a reading disorder caused by within child factors (e.g., cognitive deficits) is to assess the child using a response to intervention approach. One way in which response to intervention can be measured is through dynamic assessment.

Dynamic assessment has been shown to reduce cultural and linguistic bias (Gillam, Peña, & Miller, 1999; Gutierrez-Clellen & Peña, 2001; Lidz & Peña, 1996; Peña, 2000, 2001; Peña, Iglesias, & Lidz, 2001; Tzuriel, 2000). Dynamic assessment typically includes three distinct phases: the administration of a pretest, a teaching phase, and the administration of a posttest. This method of assessment differs from typical static assessment procedures (Lidz, 1991, 1996; Sterberg & Grigorenko, 2002). During the teaching phase, children are explicitly presented with material with the purpose of not only improving posttest scores, but to also document emerging skills and strategies that the child may be using (Haywood & Tzuriel, 2002; Peña et al., 2006). Thus, dynamic assessment does not only measure what a child presently understands (i.e., the product), dynamic assessment also measures how well a child learns and responds when presented with something new (i.e., the learning processes; Lidz, 1991). This distinction is highly relevant when assessing children from different language and/or instructional backgrounds because the impact of such contextual factors can be lessened.

Dynamic assessment can also be used as a context to assess children as they are engaged in the learning process. Examiners can make judgments about a child’s responsiveness to instruction, including the amount of teaching effort needed to facilitate learning during dynamic assessment teaching procedures. Such judgments are usually referred to as assessments of modifiability (Feuerstein & Feuerstein, 1991; Gutierrez-Clellen & Quinn, 1993; Peña, 2000; Peña et al., 2001). The concept of modifiability is foundational to underlying theories of dynamic assessment (Feuerstein & Feuerstein, 1991; Feuerstein, Falik, Rand, & Feuerstein, 2006; Lidz, 1991), and modifiability scores have been shown to yield excellent sensitivity and specificity, especially when assessing language-related skills (Gutiérrez-Clellen, Brown, Robinson-Zañartu, & Conboy, 1998; Peña, 2000; Peña et al., 2006; Peña, Quinn, & Iglesias, 1992).

A review of the literature related to dynamic assessment and its utility in predicting academic ability was conducted by Caffrey, Fuchs, and Fuchs, (2008). The researchers concluded that dynamic assessment contributed significant, unique variance to the prediction of academic achievement beyond static assessment methods. It was specifically noted that dynamic assessment was more accurate when
noncontingent feedback was used (e.g., an examiner responded to a child in a standardized manner, regardless of individual child errors), when used with students with disabilities rather than students at risk or typically achieving students, and when independent measures related to the dynamic assessment or criterion-referenced tests were used in the posttesting phase of the dynamic assessment.

The research pertaining to reading correlates have offered insight into which measures are most predictive of reading, and measures related to reading real and nonsense words are consistently highly correlated with reading ability (Hammill, 2004). When evaluating measures that are most predictive of future reading difficulty for young children who have not yet received any formal reading instruction (i.e., measures that lead to the prevention of reading difficulty), it would be easy to dismiss actual reading measures as irrelevant. However, in the context of a dynamic assessment, a skill that has not yet been acquired, such as decoding, can be taught, and a child’s response to that instruction can be measured (Fuchs, Compton, Fuchs, Bouton, & Caffrey, 2011; Fuchs et al., 2007). It would seem advantageous to capitalize on the format of dynamic assessment and to create a reading assessment with content that most closely resembles the reading construct.

A dynamic assessment of reading that uses a decoding task can contain real or nonsense words. Assessments that incorporate nonsense words have been shown to be less culturally and linguistically biased (Campbell, Dollaghan, Needleman, & Janosky, 1997; Ellis Weismer et al., 2000; Peña et al., 2006) and may be most appropriate to use with children learning English as a second language. Furthermore, in dynamic assessment, the material presented at pretest must be unfamiliar to the child, and the use of nonsense words ensures that prior exposure is unlikely.

It is also important to consider the population being assessed when evaluating the predictive validity of a dynamic assessment. Children who are culturally and linguistically diverse are usually at a disadvantage when being assessed with static measures (Demsky et al., 1998), wherein children who are not culturally and linguistically diverse will more likely produce results as expected by the test developers. Herein lies the potential of dynamic assessment and its purported ability to mitigate cultural and linguistic bias (Gutiérrez-Clellen & Peña, 2001).

Children learning to read in a second language (L2) are not homogeneous. First language (L1) background and experiences differ greatly between children, and experience with reading or reading-related skills in L1 significantly affects L2 reading (e.g., Abu-Rabia, 1997; Bialystok, McBride-Chang, & Luk, 2005; Brumian-Martin et al., 2006; Cummins, Swain, Nakajima, Handscombe, & Green, 1981; Da Fontoura & Siegel, 1995; Durgunoglu, Nagy, & Hancin, 1993; Geva & Siegel, 1999; Gholamain & Geva, 1999; Koda, 2007; Legarreta, 1979; Skutnabb-Kangas & Toukomaa, 1976; Troike, 1978; Wade-Woolley & Geva, 2000; Wang, Perfetti, & Liu, 2005). For ELLs, static assessment in L2 will not only reflect L2 reading ability, it will also reflect the experiential variance in L1 reading. Static results will be dependent upon the extent to which a child has L1 skills and can incorporate those skills through cross-linguistic transfer into L2. On the other hand, when examiner ratings of modifiability and examination of the learning process are the focus of a dynamic assessment, experiential variances in L1 are bypassed. What dynamic assessment does is shift the assessment from a focus on the construct of what a child currently knows to the construct of what a child can learn—a construct that is potentially independent of test content and prior knowledge.

Few studies of dynamic assessment of reading have focused on predicting reading ability for populations of culturally and linguistically diverse children. A review of previous dynamic assessment of reading research leads to the examination of the predictive validity of a dynamic assessment of nonsense word recoding implemented in a cross-grade longitudinal research design with culturally and linguistically diverse children. No study has yet undertaken this examination. Therefore, the purpose of this study was to determine the extent to which a dynamic assessment of reading administered at kindergarten was predictive of first-grade reading ability for bilingual Latino children and to compare the classification results of the dynamic assessment to the classification results of static reading measures.

**Method**

**Participants**

Sixty-three Latino bilingual kindergarten children who were identified as at risk for language impairment during prekindergarten screening participated in this longitudinal study. They were drawn from a larger sample of 249 children from Ogden, Utah, who participated in a multisite study of risk for language impairment in bilingual children (Bohman, Bedore, Peña, Mendez-Perez, & Gillam, 2010; Peña, Gillam, Bedore, & Bohman, 2011). The participants were screened before entering kindergarten using Bilingual English Spanish Oral Screener (BESOS; Peña, Bedore, Gutiérrez-Clellen, Iglesias, & Goldstein, 2009), which contained morphosyntax and semantics subtests in both English and Spanish. Children were considered to be at risk for language impairment if they scored below the 30th percentile on at least one subtest in each language. Thus, we used a disproportionate stratified sampling method (Koch, 1969; Sutcliffe, 1965) by preselecting Latino, bilingual (Spanish-English) students who were at risk for language impairment before entering kindergarten. By only selecting English language learners who were at risk for language impairment, we increased the probability of including a higher proportion of children who
Table 1. Participant Characteristics.

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>n</th>
<th>M</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>29 (46%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>34 (54%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hispanic/Latino</td>
<td>63 (100%)</td>
<td>65.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Age (months)</td>
<td>63</td>
<td>65.3</td>
<td>4.2</td>
</tr>
<tr>
<td>Mother’s education (years)</td>
<td>60</td>
<td>8.7</td>
<td>3.1</td>
</tr>
<tr>
<td>Father’s education (years)</td>
<td>54</td>
<td>8.1</td>
<td>3.2</td>
</tr>
<tr>
<td>Socioeconomic status (weighted Hollingshead score)</td>
<td>63</td>
<td>14.3</td>
<td>7.2</td>
</tr>
<tr>
<td>Years in preschool</td>
<td>61</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Years expressive Spanish</td>
<td>62</td>
<td>3.3</td>
<td>2.8</td>
</tr>
<tr>
<td>Years expressive English</td>
<td>62</td>
<td>0.2</td>
<td>1.1</td>
</tr>
<tr>
<td>Years expressive English/Spanish</td>
<td>62</td>
<td>2.4</td>
<td>2.7</td>
</tr>
</tbody>
</table>

Note. Data collected 1 week prior to kindergarten except for DIBELS beginning kindergarten measures, which were collected within the first 2 months of kindergarten. DIBELS = Dynamic Indicators of Basic Literacy Skills.

would have future reading difficulty (Catts & Hogan, 2003). This was expected to yield a more accurate estimate of sensitivity and specificity because these were the children who would be likely to be referred for assessment in school environments. Also, by reducing the number of participants, we were better able to efficiently administer a large battery of assessments and gather highly detailed information about the participants.

Participants were selected from three elementary schools within the Ogden School District. Across the three schools 74% to 77% of the population was Latino, 100% received free lunch, and 9% to 15% were homeless. The Ogden school district has an average student attrition rate of 33% each year (Ogden School District School Improvement Grants Application, 2010). The children attended general education classrooms in which English was the primary language of instruction.

There were 29 girls and 34 boys with a mean age of 65.3 months ($SD = 4.2$). All the participants were English language learners before entering kindergarten. Socioeconomic information, context-specific information such as English/Spanish use and exposure in the home, and language ability were collected for each subject (see Table 1). These variables were collected at kindergarten for the purpose of detailed participant description. Information obtained from the participants’ parent/guardian and school indicated that they were from low socioeconomic status backgrounds as calculated using weighted Hollingshead scores (Hollingshead, 1957), with mother and father years of education averaging 8.7 ($SD = 3.1$) and 8.1 ($SD = 3.2$) years, respectively. The majority of the children met our formal criteria for bilingualism prior to kindergarten, having at least 20% combined input and output in English and Spanish according to parent report (Peña et al., 2011). Ten percent of the children were in the emerging stages of bilingualism. These children had lived in the United States for at least 1 year, and Spanish input and output had predominated at home. Once they began their formal education in kindergarten, they experienced a significant amount of English input and output each day. Preschool attendance was variable ($M = 1.0$ years; $SD = 1.0$ years).

**Parent interview.** A language questionnaire (Gutiérrez-Clellen & Kreiter, 2003) was administered in the parent’s preferred language in person or over the telephone. This survey was developed as part of the norming process for the Bilingual English-Spanish Assessment (BESA; Peña, Gutiérrez-Clellen, Iglesias, Goldstein, & Bedore, 2009). It was designed to collect information about what activities the child participated in each hour of the day during the week and during the weekend, with whom the child interacted, and in which language the child was spoken to and responded. The parent interview was conducted just prior to the child’s enrollment in kindergarten. Parents were asked to recount a typical day during the week and during the weekend listing all of the major activities their child participated in each hour (e.g., dinner time, watching T.V.) and specifying with whom their child interacted and in which language their child was exposed to receptively and which language their child used expressively. In addition to the daily language use and exposure accounting, the parents were asked to summarize their child’s language use at home and at preschool from birth to the present-day. This yearly expressive language information was collected to better understand the extent that the child used English, Spanish, or both languages across their life span.

Using yearly language use data from the language questionnaire collected prior to kindergarten, the parents were classified into five language groups following procedures similar to those used by Peña et al., (2011); Hammer, Miccio, and Rodriguez (2004); and Kohnert, Bates, and Hernández (1999). The five language groups were as follows: functional monolingual English (i.e., 80% of expressive language history was English), functional monolingual Spanish (i.e., 80% expressive language history was Spanish), bilingual English dominant (60%−80% expressive language history was English), bilingual Spanish dominant (60%−80% expressive language history was Spanish), and balanced bilingual (40%−60% expressive language history was English and Spanish). None of the children were...
functional monolingual English, 46% (n = 29) of the participants were functional monolingual Spanish, 16% (n = 10) were bilingual Spanish dominant, 3% (n = 2) were bilingual English dominant, and 38% (n = 24) were balanced bilingual. These data were collected 3 months prior to administration of the predictor measures and likely underestimate the participants’ English language proficiency.

BESA. The full version of the Bilingual English-Spanish Assessment (Peña, Gutierrez-Clellen, et al., 2009) was used to assess the language abilities of the children. The BESA consists of four tests (morpho-syntax, semantics, phonology, and pragmatics) in Spanish and in English. The format of the tests in each language is similar, but the test items are representative of critical aspects of speech and language development in the target language. Across the four subtests in Spanish, coefficient alpha values range from .784 to .840. Coefficient alpha values for the four English subtests range from .812 to .918, indicating good internal consistency within subtests for both languages.

Information from the parent questionnaire, which yielded language dominance classifications, and information from the BESA subtests, which yielded classifications of average or below average language ability in English and Spanish, were cross-referenced to depict a more accurate account of language ability. Those children who were Spanish language dominant according to the language profile questionnaire, and who scored below the 30th percentile on the Spanish semantics and morphosyntax subtests of the BESA were classified as having Spanish language difficulty. Likewise, children who were classified as English language dominant and who scored below the 30th percentile on the English semantics and morphosyntax subtests were classified as having English language difficulty. None of the children who were classified as bilingual scored below the 30th percentile on all four English and Spanish BESA semantics and morphosyntax subtests. Thus, the results of the BESA in combination with the language profile information resulted in two different language ability groups: those having typical language ability (TL) and those having language difficulty (LD). These classifications are considered to be a gross estimate of language ability.

 Measures

Dynamic assessment of nonsense word decoding. The dynamic assessment consisted of three phases: (a) the pretest phase, where the examiner asked the children to recode four nonsense words; (b) the teaching phase, where the children participated in a brief lesson on how to recode the nonsense words using an onset-rime, analogous strategy in conjunction with whole word recognition; and (c) a posttest phase, where the examiner asked the children to recode the same nonsense words displayed in a different order. The pretest (and posttest) contained words that followed a consonant-vowel-consonant pattern, with the short vowel /ae/ and the final consonant /d/ consistent in each word, thereby creating rhyming words (i.e., tad, zad, nad, kad). Because English was the primary language of reading instruction for these participants, the dynamic assessment included words that related closely to the English language. Thus, the four nonsense words were selected to represent phonotactically and orthographically typical patterns encountered in English, with the expectation that the test content would predict future reading ability in English.

Therefore an analogous decoding strategy was taught during the teaching phase of the dynamic assessment because of its immediate application to the four nonsense words. Although another approach to teaching the nonsense words could have been applied (e.g., phonological analysis), the analogous approach was the most logical given that we wanted to teach the children how to read the four specific nonsense words that rhymed in the least amount of time necessary. By taking advantage of the onset-rime structure of each word, the analogous approach was the preferred choice.

Children must be unsuccessful during the pretest phase of the dynamic assessment so that there is opportunity to learn during the teaching phase. Therefore, if the children were able to recode 75% of the nonsense words or recode 75% of the phonemes within the nonsense words at pretest, the test was discontinued. This discontinuation was further justified by reasoning that if a child at the beginning of kindergarten was able to recode at least three out of four nonsense words, then the likelihood of that child being at risk for future reading difficulty was quite small. A total of 35% of the participants (n = 22) met the criteria for termination of the dynamic assessment at pretest.

After the examiner administered the dynamic assessment pretest, the children were told that they were going to be taught how to read, and the examiner gave a brief lesson on how to recode the pretest words. The teaching phase of the dynamic assessment used a noncontingent feedback approach that lasted less than 3 minutes. The examiner pointed to an individual grapheme (onset) and (using the nonsense word tad as an example) said, “This letter says /t/. “Say /t/.” The examiner then pointed to the graphemic cluster (rime) and said, “These letters together say /æd/.” “Say /æd/.” The examiner then said, “Put them together and you get the word t-ad, tad. What is this word?” The examiner then successively taught the child to blend the onset-rime, graphophonemic information to recode each word (i.e., tad, zad, nad, kad).

A second round of teaching immediately followed the first round, where the examiner asked the child to imitate his or her reading. This second round of teaching was designed to reduce the level of examiner instruction and facilitate independent recoding and word recognition. To
illustrate, if an examiner was teaching a child to recode the word *tad* during the second round of instruction, the examiner would say */t/* while pointing to the letter *t*. The child would imitate this model. Then the examiner followed the same procedures by pointing to the letters *ad*, the onset and rime *t-aed*, and the whole word *tad*, with the child imitating the examiner each time. If the child independently produced the targets, then he or she was allowed to continue in that manner. If the child did not independently produce the targets, he or she was prompted to do so.

The posttest phase immediately followed the teaching phase of the dynamic assessment. During the posttest phase, the examiner showed the children the same words from the pretest and teaching phases, with the words presented in a different order. The children were then asked to read the words. During the posttest the examiner was not allowed to help the child; however, some relatively neutral prompts were allowed such as “*You can guess—what do you think this word says?*” or “*Remember what I told you?*” These prompts were occasionally necessary when children were particularly reticent and did not want to perform independently. The individual phonemes and words recorded correctly during the pretest and posttest were recorded using the international phonetic alphabet (IPA).

**Dynamic assessment phoneme gain score.** A simple approach to scoring the dynamic assessment was to subtract the number of phonemes in the nonsense words read correctly at pretest from the number of phonemes read correctly at posttest. This gain score approach, while desirable because of ease of calculation, offered limited information. If we had used a simple gain score measure, students who already had experience with reading would be at a disadvantage because they would have less information to learn from pretest to posttest. For example, if a student read 6 out of 12 phonemes correctly on the dynamic assessment pretest, he or she would have fewer phonemes to learn than the student who was unable to read any phonemes at pretest. If, after the intervention phase of the dynamic assessment, those same two students read all 12 phonemes correctly at posttest, then the student with the lower pretest score (0) would get the higher gain score (posttest 12 – pretest 0 = 12), and the student who had a higher pretest score (6) would get a lower gain score (posttest 12 – pretest 6 = 6). Therefore, this simple gain score approach would not necessarily reflect a student’s response to instruction and would penalize those students with greater initial reading ability. Nevertheless, for exploratory purposes, a simple gain score of sounds was calculated because of its traditional approach to scoring nonsense word assessments.

**Dynamic assessment residuum gain score.** Dynamic assessment is the most useful for children who do not perform a particular task at an age-appropriate level (Peña et al., 2006). Children who have the most to gain (the greatest difference between their pretest performance and the ceiling point of a particular task) have the potential to show larger gains after the teaching phase of dynamic assessment (Kester, Peña, & Gillam, 2001). In an attempt to account for prior knowledge, we calculated a percentage of residuum, which was the percentage of the remainder between the pretest score and the ceiling. The dynamic assessment residuum gain score represented the relative amount of information yet to be learned after taking into account performance on the pretest.

This process can be illustrated by referring back to the two students in the previous example. Recall that the student who correctly read 6 out of 12 sounds during the pretest had 6 more sounds to learn before posttest mastery. After the intervention phase of the assessment, that student read all 12 sounds correctly at posttest. By using a relative scoring procedure, the student had 6 more sounds to learn, and that student learned all 6 sounds, resulting in a percentage residuum gain score of 100% (6 out of 6). The student who read 0 sounds correctly at pretest and then read 12 sounds correctly at posttest also received a percent residuum gain score of 100%. Both students responded perfectly to the intervention, and both students received a percentage residuum gain score of 100%.

The dynamic assessment residuum gain score was designed to account for individual sounds learned and individual words learned. Because the examiner provided noncontingent feedback during the teaching phase, the focus of the scoring was based on each child’s performance given standardized instruction. The intensity and length of instruction were held constant.

**Decoding strategy score.** The dynamic assessment was also designed to assess the participants’ ability to use an analogous decoding strategy during posttest. The ability to apply an analogous decoding strategy or to quickly form mental orthographic representations (Apel, 2009; Wolter & Apel, 2005), the principle focus of the teaching phase of the dynamic assessment, was assessed by analyzing reading strategies used by the participants during the posttest of the dynamic assessment. A scoring rubric was used to calculate the reading strategy score (Appendix A). We assigned different, hierarchical values to (a) a whole word guessing strategy that implicated erroneous word recognition, (b) a phonological analysis strategy that had a specific focus on the initial letter of each nonsense word, (c) an analogous decoding strategy, and (d) a whole word recognition strategy that implicated the use of accurate mental orthographic representations (Laing, 2002; McGuinness, 1997).

Whole word guessing was defined as the production of one or more words that contained two or more incorrect sounds (e.g., a child produced */tap/* for */kad*/) or when two
or more words were repeated (e.g., a child produced /kad/ each time when attempting to recode /rad/, /nad/, and /zad/). Evidence that the participant learned to use an onset, phonological analysis decoding strategy was defined by an increase in the accurate decoding of an initial letter-sound from pretest to posttest (e.g., a child at pretest produced /gat/ for the nonsense word /kad/ but at posttest the child produced /kat/ for the nonsense word /kad/). Evidence that the participant learned to use an analogous decoding strategy was defined by the use of an onset-rime production at posttest (e.g., a child produced /t/ and the rime cluster /ad/ to recode the nonsense word /tad/). Evidence that the participant used a whole word recognition strategy was defined by the rapid, accurate recoding of the nonsense word.

The strategy scoring procedure was comprised of the following guidelines: Participants received 1 point for using a whole word guessing strategy and when no analogous strategy was applied and no improvement in initial phonological analysis strategy was evident. Participants received 2 points when whole word guessing was evident and there was no analogous decoding strategy applied (but there was evidence of improvement in initial phonological analysis). Participants received 3 points when they used a combination of whole word guessing and an analogous decoding strategy. Participants received 4 points when whole word guessing was not evident.

**Response to instruction score.** In addition to the decoding strategy analysis, each student was rated on a 3-point response to instruction scale based on each examiner’s perception of how difficult it was for the participant to respond to the instruction, ranging from easy to difficult (Appendix A). Students who had a high degree of difficulty in responding to the standardized instructions of the teaching phase of the dynamic assessment were assigned a response to instruction score of 0. Those students whose responsiveness was moderate received a score of 1, and those students whose responsiveness was considered easy received a score of 2. These response to instruction ratings were based on the judgment of the examiner at the time of the testing procedure.

**Dynamic assessment modifiability score.** The decoding strategy score (0–4) and response to instruction score (0–2) were combined to create a dynamic assessment modifiability score, which reflected the extent to which each child was considered responsive by the examiner during the teaching phase of the dynamic assessment and the extent to which the child applied the reading strategy taught during the teaching phase to the recoding task at posttest (Appendix A).

**Administration fidelity and scoring reliability.** Administration fidelity and scoring reliability for the dynamic assessment were analyzed in a follow-up replication study. The identical standardized, dynamic assessment of reading, with an analogy teaching strategy and a modifiability rating scale, was administered to a large sample of kindergarten children at the beginning of the school year by five independent examiners who had the same level of education (first-year graduate students) and who received the same level of training as the examinees in the current study. During data collection, the first author and a certified speech-language pathologist observed 59 of the dynamic assessment administrations and used a fidelity checklist to document the extent to which administration and IPA transcription were performed consistently. Procedures for administration and transcription were completed correctly in 97% (75%–100%) of the observed sessions.

Examiners scored the total number of correct sounds, total number of correct words, and completed ratings for the response to instruction during the administration of the dynamic assessment. During the administration of the dynamic assessment, the first author and a certified speech-language pathologist independently scored the total number of correct sounds, total number of correct words, and completed the responsiveness scale (Appendix A). Interrater reliability of total number of correct sounds, words, and responsiveness was 97% (75%–100%). Two research assistants scored children’s strategy use after data collection was completed using the IPA transcriptions and the strategy scoring rubric (Appendix A). Thirty dynamic assessments were randomly chosen and scored by both research assistants. Interrater reliability ranged from 75% to 100%, with a mean of 98%.

**Static kindergarten reading measures.** The Ogden City School District administers the English *Dynamic Indicators of Basic Literacy Skills* (DIBELS; Kaminski & Good, 1996) to all students in every elementary school grade three times per year. The DIBELS assessment is administered in English and is used to gauge progress at the student, school, and district levels. Different DIBELS measures are administered depending on whether it is the beginning, middle, or end of the school year and depending on the grade level of the student. At the beginning of the participants’ kindergarten school year, the school district administered the *Initial Sound Fluency* subtest (ISF) and the *Letter Naming Fluency* subtest (LNF). The *Initial Sound Fluency* measure is a phonological awareness task that requires a child to recognize and produce the initial sound of as many words as possible that are presented orally in 1 minute. The *Letter Naming Fluency* measure requires that children name as many upper- and lower-case letters arranged in random order as fast as possible in 1 minute.

**Static first-grade reading measures.** The DIBELS (Kaminski & Good, 1996) *Nonsense Word Fluency* (NWF) subtest was the measure selected to represent decoding ability. The DIBELS NWF is a standardized assessment designed to
record the number of nonsense words a child can produce in 1 minute. The test is comprised of randomly ordered VC and CVC nonsense words (e.g., rav, sig, ov) and is scored according to the number of correct sounds produced. The test-retest reliability was reported to be .83 (Good et al., 2004). The convergent-corrallational evidence of validity of the DIBELS NWF when compared to a curriculum-based measure of reading fluency administered 4 months later was .82 (Good et al., 2004).

The DIBELS Oral Reading Fluency subtest was used to represent reading fluency. The Oral Reading Fluency subtest required that children read a passage aloud for 1 minute. Words substituted, omitted, and any hesitations that lasted longer than 3 seconds were scored as errors. The total number of words read correctly in 1 minute comprised the oral reading fluency score. The test-retest reliability was reported to range from .92 to .97, and the alternate form reliability of different passages taken from the same level ranged from .89 to .94 (Tindal, Marston, & Deno, 1983). Convergent-corrallational evidence of validity measured across eight different studies ranged from .52 to .91 (Good & Jefferey, 1998).

The Woodcock Reading Mastery Test–Revised (WRMT-R; Woodcock, 1987) Word Identification subtest (Word ID) was used to represent word identification. The Letter–Word Identification subtest involved the recognition of sight words that were presented in isolation. The split-half reliability for first grade Word Identification was .98. The internal consistency reliability (Cronbach’s alpha) was reported to be .96. Convergent-corrallational evidence of validity with the Woodcock-Johnson Reading Test was .69.

### Procedures

During Phase 1 testing, 63 Latino kindergarten children were administered the nonsense word recoding dynamic assessment by our research team and were administered DIBELS subtests by the school district personnel. Phase 2 testing took place a little less than 2 years after the kindergarten dynamic assessment was administered, near the end of the participants’ first-grade school year. This phase included the collection of school district–administered measures pertaining to nonsense word fluency and reading fluency. Children also received a standardized, norm-referenced assessment of word identification. The three first-grade English reading measures (nonsense word fluency, oral reading fluency, and word identification) were chosen to serve as formative criterion measures representative of the construct of word-level reading (Kamhi, 2009). At the end of first grade, 76% (n = 48) of the original kindergarten participants were available for testing and were administered the DIBELS nonsense word fluency and oral reading fluency subtests. This degree of attrition was not surprising due to the high attrition rate for the school district. One week later the word identification subtest was administered. Sixty percent (n = 38) of the participants were available for testing on that day.

### Results

**Descriptive Statistics**

Table 2 summarizes descriptive statistics for the kindergarten dynamic assessment predictor measures, the DIBELS middle kindergarten measures, and the first-grade criterion measures. Intercorrelations among language ability, the kindergarten measures, and the first-grade measures are shown in Table 3. Intercorrelations among language ability, the dynamic assessment predictor measures, and the first-grade word-level reading measures indicated nonsignificant results. The three different ways in which the dynamic assessment was scored (sounds gain, residuum gain, and modifiability score) produced different correlation results, with the dynamic assessment modifiability score rendering the highest correlations with the three first-grade reading outcome measures, which were statistically significant. The dynamic assessment residuum gain score was also significantly correlated with the three first-grade reading measures, whereas the dynamic assessment sounds gain score was not significantly correlated with any of the three outcome measures.

### Table 2. Descriptive Statistics of Kindergarten Dynamic Assessment Predictor Measures and First-Grade Reading Outcome Measures.

<table>
<thead>
<tr>
<th>Measure</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kindergarten dynamic assessment</td>
<td>63</td>
<td>6.7</td>
<td>4.6</td>
<td>7.0</td>
</tr>
<tr>
<td>Dynamic assessment sounds gain</td>
<td>63</td>
<td>58%</td>
<td>37%</td>
<td>56%</td>
</tr>
<tr>
<td>Dynamic assessment residuum gain</td>
<td>63</td>
<td>42.4</td>
<td>19.0</td>
<td>50.0</td>
</tr>
<tr>
<td>Dynamic assessment modifiability score</td>
<td>63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kindergarten DIBELS</td>
<td>60</td>
<td>6.1</td>
<td>6.6</td>
<td>0.0</td>
</tr>
<tr>
<td>Initial sound fluency</td>
<td>60</td>
<td>3.8</td>
<td>7.3</td>
<td>0.0</td>
</tr>
<tr>
<td>Letter naming fluency</td>
<td>48</td>
<td>53.9</td>
<td>17.9</td>
<td>51.0</td>
</tr>
<tr>
<td>Nonsense word fluency</td>
<td>48</td>
<td>19.7</td>
<td>12.9</td>
<td>16.0</td>
</tr>
<tr>
<td>Oral reading fluency</td>
<td>48</td>
<td>39.2</td>
<td>12.1</td>
<td>40.5</td>
</tr>
<tr>
<td>Word identification</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note. Dynamic assessment sounds gain, dynamic assessment residuum gain, and dynamic assessment modifiability score (the combined dynamic assessment decoding strategy score and the dynamic assessment response to instruction score) represent the three approaches to scoring the dynamic assessment. Initial sound fluency, letter naming fluency = DIBELS beginning of kindergarten year measures; Nonsense word fluency and oral reading fluency = DIBELS end–of–year first–grade measures; word identification = WRMT-R first-grade word identification subtest; DIBELS = Dynamic Indicators of Basic Literacy Skills; WRMT-R = Woodcock Reading Mastery Test–Revised.
Assumptions of Linear Multiple Regression

Regression analysis requires that several assumptions be met before the sample-derived results can be projected inferentially to a population. Visual and statistical analysis of the data (in most cases the residuals) can reveal problems with the regression model. For this study, some of the assumptions required for regression analysis were examined in a secondary manner via the extensive visual and statistical analysis of other assumptions. The assumption that there was a linear relationship between all of the Y variables (criterion measures) and the X variables (predictor measures) was assessed in part through the examination of other assumptions (e.g., detection of outliers and homoscedasticity). Had any of the analyses indicated that the relationship between X and Y was nonlinear, we were prepared to use a weighted least squares (WLS) regression analysis or transform the data to fit a linear regression model to simplify the relationship. This was not necessary.

Outliers. It is possible for outliers to have a powerful effect on the results of regression analysis. Regression analysis results can reflect a small number of atypical cases instead of the general data trend. If a sample size is small, the potential for outliers to affect the regression analysis increases. High outlying data were winsorized to the next highest data point +1, and low outlying data were winsorized to the next lowest data point −1. All data points identified as outliers, whether identified as having high discrepancy or not, were also winsorized in the same manner.

Normality. A normal distribution of scores meets one of several important assumptions required for the inferential application of regression analysis (Cohen, Cohen, West, & Aiken, 2003). After winsorizing the extreme outliers from NWF, ORF, and Word ID, visual examination of the data via histograms and Q-Q plots revealed that the distribution of the raw scores within NWF, ORF, and Word ID score appeared to follow a relatively normal distribution. Figure 1 shows the histograms of the NWF, ORF, Word ID data, and Figure 2 shows the Q-Q plots of the NWF, ORF, Word ID data.

Linearity. The Levene test of homogeneity of variance (Levene, 1960) indicated that the variance of the residuals was constant with the exception of the regression equation that included the dynamic assessment strategy measure and the oral reading fluency criterion measure ($p = .01$). To determine whether corrective action was necessary (e.g., data transformation or WLS regression), an estimate of the magnitude of the nonconstant variance was calculated. Unless the magnitude of the heteroscedasticity is large, the nonconstant variance will not have a material effect on the results of the regression analysis (Cohen et al., 2003). The magnitude of the nonconstant variance was determined by ordering the cases from lowest to highest according to their values on the predictor variable (dynamic assessment strategy) and then by dividing those cases into three sets (slices) with an equal number of cases in each slice.

The variance of the residuals around the regression line within each slice was calculated, revealing a mean square error of 120.9, 172.3, and 123.0 for each slice, respectively. It was determined that if the ratio of the largest to the smallest conditional variance for the slices exceeded 10, or if the conditional variance changed in a regular and systematic manner as values on the predictor variable (dynamic assessment strategy) and then by dividing those cases into three sets (slices) with an equal number of cases in each slice.

The variance of the residuals around the regression line within each slice was calculated, revealing a mean square error of 120.9, 172.3, and 123.0 for each slice, respectively. It was determined that if the ratio of the largest to the smallest conditional variance for the slices exceeded 10, or if the conditional variance changed in a regular and systematic manner as values on the predictor variable (dynamic assessment strategy) and then by dividing those cases into three sets (slices) with an equal number of cases in each slice.

Multiple Regression Analysis

A preliminary regression analysis was conducted to determine whether the dynamic assessment residuum gain score uniquely accounted for variance in the criterion variables over and above the dynamic assessment modifiability score. This regression analysis informed whether the two different
dynamic assessment measures (previously identified as being significantly correlated with the criterion measures) should be included in the regression model or whether the dynamic assessment modifiability score alone should be included. The results of the preliminary regression analysis indicated that the dynamic assessment residuum gain score did not account for a significant amount of unique variance within NWF, $R^2$ change = .01, $F(1, 45) = .40, p = .53$; ORF, $R^2$ change = .004, $F(1, 45) = .23, p = .64$; or Word ID, $R^2$ change = .001, $F(1, 35) = .07, p = .80$.

Based on the preliminary analysis, a multiple regression analysis was conducted to assess the extent to which the dynamic assessment modifiability score accounted for a significant amount of variance in word-level reading at first grade. Results indicated that the dynamic assessment was significantly related to the three different criterion measures. The dynamic assessment accounted for 19% of the variance of first-grade nonsense word fluency, $R = .46, R^2 = .21$, adjusted $R^2 = .19, F(1, 46) = 8.82, p < .01$; 19% of the variance for first-grade oral reading fluency, $R = .46, R^2 = .21$,

Figure 1. Histograms representing the distribution of nonsense word fluency, oral reading fluency, and word identification raw scores.
adjusted $R^2 = .19, F(1, 46) = 12.19, p < .001$; and 24% of the variance for first-grade Word ID, $R = .51$, $R^2 = .26$, adjusted $R^2 = .24$, $F(1, 36) = 12.38, p < .001$. Results of the regression analysis are detailed in Table 4.

**Classification Analysis**

The main purpose of the kindergarten dynamic assessment measure was to accurately predict which bilingual Latino children would be at risk for reading difficulty at the end of first grade. A classification analysis as described by Lichtenstein and Ireton (1984) was used to determine the sensitivity and specificity of the measures. Kindergarten children were classified as at risk for reading difficulty if they met the at-risk criterion on the dynamic assessment modifiability score, which was a score of 2 or lower (see Appendix A). The dynamic assessment modifiability score was the sum of the decoding strategy score (0–4) and the response to instruction score (0–2). Thus, a score of 2 or lower indicted that the participant either scored a 2 on the decoding strategy scale (which indicated whole word guessing and no evidence of onset-rime strategy use) and a 0 on the response to instruction scale (indicating that the examiner considered it difficult for the child to respond to the instruction) or a 1 on the decoding strategy scale (the lowest score possible) and a 1 on the response to instruction scale (indicating a moderate child response). At-risk reader status in first grade was determined by scoring at or below the 20th school district percentile on the NWF or ORF measure or at or below the 20th percentile on the WRMT-R word identification measure, using the test’s published norms.
Participants who were at or below criterion on the kindergarten dynamic assessment predictor measure and also at or below the 20th percentile criterion on one or more of the reading measures at first grade were classified as true positives. Participants who performed above the criteria cutoff values on both the kindergarten and first-grade measures were classified as true negatives. False positives included students who met the kindergarten criterion but were not identified on first-grade measures as having reading difficulty, whereas false negatives included students who did not meet the kindergarten criterion but scored at or below the 20th percentile on the first-grade reading measure.

Sensitivity, the ability of a kindergarten measure to identify those first-grade students who scored at or below the 20th percentile on a reading measure, was calculated by dividing the number of true positives by the sum of the true positives and false negatives. Specificity, the ability of the kindergarten measure to identify students who were reading above the 20th percentile on a reading measure, was calculated by dividing the number of true negatives by the sum of the true negatives and false positives. The results are presented in Table 5.

Near the beginning of kindergarten, the dynamic assessment identified 25% of the participants as at risk for future reading difficulty (16/63). The classification analyses indicated that the kindergarten dynamic assessment modifiability score was perfectly sensitive (demonstrating 100% sensitivity) for identifying those participants who would score at or below the 20th percentile on the first-grade ORF measure and the first-grade Word ID measure. This level of sensitivity is meaningful when considered in conjunction with specificity, which was 80% and 88%, respectively. The dynamic assessment modifiability score also resulted in high levels of sensitivity (86%) and specificity (85%) for NWF.

Because the dynamic assessment modifiability score was the combination of the dynamic assessment decoding strategy score and the dynamic assessment response to instruction score, we were interested in examining the extent to which dynamic assessment accurately classified the participants using only the objective scoring procedure (the decoding strategy score). Thus, we used a cutoff score of 2 for the decoding strategy score and calculated the sensitivity and specificity of that measure for each of the first-grade formative criterion measures. Results indicated that the dynamic assessment decoding strategy score yielded 100% sensitivity and 62% specificity for ORF, 100% sensitivity and 68% specificity for Word ID, and 86% sensitivity and 76% specificity for NWF. Therefore, the response to instruction ratings were an informative aspect of the dynamic assessment procedure.

To explore the classification accuracy of the static kindergarten DIBELS measures, and to compare those results to our dynamic assessment findings, we first examined the percentage of participants who were initially classified at risk in kindergarten by DIBELS (Table 1). Results indicated that over 95% (60/63) of the participants were classified as needing intensive or strategic services near the beginning of kindergarten, with nearly 50% (31/63) classified as needing

### Table 4. Multiple Regression Analysis Estimating the Unique Variance Associated With the Dynamic Assessment Using the First-Grade Reading Outcome Measures.

<table>
<thead>
<tr>
<th>First-Grade Reading Outcome Measure</th>
<th>B</th>
<th>SE</th>
<th>Beta</th>
<th>p</th>
<th>Adjusted R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsense word fluency (n = 48)</td>
<td>.44</td>
<td>.12</td>
<td>.46**</td>
<td>&lt;.01</td>
<td>.19</td>
</tr>
<tr>
<td>Oral reading fluency (n = 48)</td>
<td>.31</td>
<td>.09</td>
<td>.46**</td>
<td>&lt;.01</td>
<td>.19</td>
</tr>
<tr>
<td>Word identification (n = 38)</td>
<td>.33</td>
<td>.09</td>
<td>.51***</td>
<td>&lt;.01</td>
<td>.24</td>
</tr>
</tbody>
</table>

Note. All analyses conducted with raw scores. The dynamic assessment modifiability score (the combined dynamic assessment decoding strategy score and the dynamic assessment response to instruction score) was used as the predictor measure.

### Table 5. Classification Analysis.

<table>
<thead>
<tr>
<th>DA modifiability score</th>
<th>NWF Sensitivity</th>
<th>NWF Specificity</th>
<th>ORF Sensitivity</th>
<th>ORF Specificity</th>
<th>Word ID Sensitivity</th>
<th>Word ID Specificity</th>
</tr>
</thead>
</table>

Note. DA modifiability score = dynamic assessment modifiability score (the combined dynamic assessment decoding strategy score and the dynamic assessment response to instruction score); NWF = DIBELS first-grade nonsense word fluency; ORF = DIBELS first-grade oral reading fluency; Word ID = WRMT-R first-grade word identification subtest; DIBELS = Dynamic Indicators of Basic Literacy Skills; WRMT-R = Woodcock Reading Mastery Test–Revised.
intensive services. This high percentage immediately eliminates the need for further analysis for predictive accuracy. When a measure classifies nearly all children as being at risk, the inevitable result will be near-perfect sensitivity with alternating near-imperfect specificity.

**Discussion**

This study was designed to assess the predictive validity of a dynamic assessment of decoding administered to bilingual Latino kindergarteners who were at risk for language impairment. Validity was viewed as a unitary concept evidenced by carefully defined constructs (Messick, 1993) and the assessment of those constructs, which yield appropriate inferences and actions (Messick, 1989). In this case, the narrow view of reading served as the theoretical foundation of the predictor and criterion measures. We hypothesized that a dynamic assessment predictor variable would be related to a static assessment criterion variable if both were conceptually driven by the narrow view of reading.

We developed three predictor indices consequent of the dynamic assessment of decoding: a dynamic assessment raw gain, a dynamic assessment residuum gain, and a dynamic assessment modifiability score, which reflected both an examiner judgment of response to instruction and an analysis of the reading strategy applied at posttest. The dynamic assessment predictor measure was administered to bilingual kindergarteners, and the criterion measures were administered to the same children during first grade. All the children in the study performed in the at-risk range on English and Spanish language measures administered before kindergarten. We included this subgroup of kindergarten children because they would typically be the children who would be of concern to educators during their kindergarten year and because it increased the likelihood of including children with reading difficulties in the sample.

**Multiple Regression**

Multiple regression analysis was used to assess content-relevance validity because it provided evidence of a significant linear relationship between predictor and criterion measures together with estimates of effect sizes. The regression analysis indicated that the dynamic assessment modifiability score significantly predicted first-grade nonsense word fluency, oral reading fluency, and word identification.

**Predictive Validity**

The preponderant verification of construct validity was evidenced inferentially in the measures’ classification accuracy. The classification analyses indicated that the measures of dynamic assessment modifiability had excellent predictive evidence of validity. The dynamic assessment modifiability score yielded 100% sensitivity and 80% specificity for first-grade oral reading fluency, 100% sensitivity and 88% specificity for first-grade word identification, and 86% sensitivity and 85% specificity for first-grade nonsense word fluency. These ranges of sensitivity and specificity meet or exceed minimal acceptable levels (Plante & Vance, 1994), and we consider higher sensitivity over specificity to be the desired outcome of a screening instrument.

Static measures have a long-standing history of not yielding adequate classification accuracy. As previously reviewed, no research has reported adequate sensitivity or specificity results using static reading measures with bilingual children. Our findings align with previous research findings. Specifically, DIBELS static measures at the beginning of kindergarten classified almost every one of our participants (95%) as being at risk for reading difficulty, which inevitably led to inaccurate classification (Rodgers, 1983; Shaywitz, Escobar, Shaywitz, Fletcher, & Makuch, 1992; Silva, McGee, & Williams, 1985).

To our knowledge, this is the first prospective, across-grade, early reading study of bilingual children to report such high levels of sensitivity while maintaining acceptable specificity. Furthermore, classification resulting in 100% sensitivity with 80% or higher specificity was replicated across three first-grade criterion measures of distinct design. This finding is remarkable given that the dynamic assessment procedure required less than 5 minutes to administer and was only administered in English.

There are several possible reasons why the dynamic assessment used in this study resulted in such high classification accuracy 1 year later. Previous research has indicated that measures closely resembling authentic reading tasks were most predictive of reading ability for the general population (Hammill, 2004). Using actual reading tasks in an assessment designed to identify children at risk for reading problems before reading difficulty emerged would be impossible using static measures. However, because of the test-teach-retest arrangement of the dynamic assessment process, measures that children are not yet capable of doing without assistance can be a focus of investigation. Thus, the successful classification results derived from the dynamic assessment may be attributable to its suitability in measuring actual reading ability when standardized guidance is proffered.

The examination of what was modified likely contributed to the classification accuracy of the dynamic assessment. The dynamic assessment phoneme gain score, reflecting of the raw score gain in phoneme accuracy from pretest to posttest, was not significantly predictive of the first-grade criterion measures. When a modification of that gain score was calculated using a percentage residuum analysis, the distance between pretest performance and the ceiling of the posttest measure, the dynamic assessment was
significantly correlated with each criterion measure. This result is consistent with other studies showing that dynamic assessment is most revealing of learning potential in children who have difficulties performing a task. Children who perform a task well have less room for improvement after the teaching phase in dynamic assessment in comparison to children who perform a task poorly.

When the dynamic assessment measure was scored by analyzing in greater detail the response to instruction rendered during the teaching phase, and the scores assigned to different reading strategies evidenced by the participants were differentially weighted to reflect those reading strategies used most frequently by proficient readers, the dynamic assessment measure correlated to a moderately high to high degree with the first-grade criterion measures and yielded excellent classification results. Hence it was the examiners’ judgment of the children’s responsiveness to the instruction and the degree to which the participants applied the instructions given during the teaching phase to the posttest that was most predictive of future reading ability. It is interesting to note that when the examiners’ estimation of responsiveness was removed from the dynamic assessment modifiability score, specificity was reduced but sensitivity was not affected. Just by using the more objective decoding strategy score, the dynamic assessment maintained excellent sensitivity, with specificity ranging from 62% to 76%. This suggests that many children who were at risk for reading decoding problems in first grade failed to use the strategies they were taught during the posttest of dynamic assessment when they were kindergarteners. However, that measure alone should not be considered to be a diagnostic marker because some children who failed to apply the strategies they were taught during dynamic assessment turned out to have decoding abilities that were within the normal limits in first grade. The combination of strategy use at posttest and responsiveness during the dynamic assessment teaching phase was the best predictor of performance below or above the at-risk range during first grade.

Factors such as SES, L1 and L2 proficiency, and prior formal and informal reading instruction can greatly impact static L2 assessment results and the interpretation of those results. SES, prior instruction, and L2 proficiency are factors that many would consider to be logically tied to L2 reading outcomes. However, L1 proficiency and its role in L2 reading development are often overlooked. Cross-linguistic transfer from L1 to L2 has considerable influence on L2 reading proficiency. When a static measure in L2 is administered to an ELL child, the results of that assessment will be biased by the child’s L1 proficiency and L1 reading foundation. Dynamic assessment, because of the construct that it is assessing, has been posited as a means to control for the influence of all of these factors, including L1 proficiency. As previously noted, dynamic assessment allows for the measurement of the construct of modifiability. Modifiability is presumed to be an underlying, within-child trait that relates to learning potential. If measured appropriately, this learning potential can be assessed without many confounding factors biasing the results. The results of this study indicated that the participants’ L1 or L2 ability at the beginning of kindergarten did not relate to the kindergarten dynamic assessment results. This finding supported the independent nature of the construct of modifiability. Interestingly, L1 and L2 ability were not significantly correlated with any of the first-grade reading outcomes. This finding was likely due to the fact that we confined our assessment to word-level reading at first grade. If we had expanded our first-grade construct to include reading comprehension, L1 and L2 would have likely had a significant relationship with that outcome. Language ability is clearly connected to reading comprehension.

Although the dynamic assessment was able to circumvent many of the factors that often bias static assessment results, this is not to say that such factors should be ignored when teaching ELLs to read in L2. For ELLs, dual-language involvement is the defining characteristic of L2 reading (Koda, 2007). L2 reading has an inextricable connection to L1 proficiency. Bypassing these linguistic factors in the assessment process to identify a true “within-child” deficit makes sense. On the other hand, acknowledging, identifying, and taking advantage of these factors during instruction would be important for successful L2 reading acquisition.

**Limitations**

Readers should consider that the analyses in this study were exploratory, with a small sample of Latino children who were ELLs and who were at risk for language impairment. The results suggest that replication across a different, larger sample of Latino ELLs is warranted. Similar research with different groups of children other than Latino ELLs and children at risk for language impairment should also take place. Because an assessment procedure appears to have evidence of predictive validity for one group of children does not mean that it has universal applicability across all cultures and languages. In addition, there was 24% attrition in the subject pool, with 48 of 63 participants who were tested at kindergarten available for testing at the end of first grade. Only 38 of the participants were available for the word identification testing. This attrition rate, although expected due to the school district attrition rate of 33%, limits our findings, especially because of the relatively small initial sample size. To add confidence to our findings, we reviewed the profiles of the children who were not available to test at the end of first grade and compared them to those students who were available to test. The children who were no longer in the school district at the end of first grade (attrition group) matched the remaining sample of students (longitudinal group) across gender, age, SES, and kindergarten
dynamic assessment performance (see Table 6). That is, their distribution across those variables was similar to the distribution of the remaining participants, indicating that those students who did leave the district represented a random sample of the initial subject pool.

The word identification data should be interpreted with caution because of the limited number of students available to test for that measure at the end of first grade. Even so, it is important to note that all three of the outcome measures yielded high classification accuracy and that even if word identification were discarded as a criterion measure, the remaining two criterion measures (NWF and ORF) provide sufficient evidence to merit future research into the predictive validity of the dynamic assessment used in this study with a larger sample of children.

The report of fidelity of administration of the dynamic assessment and interrater reliability was derived from an ongoing follow-up study that replicated the procedures used in this study. There are no direct data on fidelity and reliability for the examiners who specifically administered and scored the dynamic assessment for this study. However, the highly standardized procedures did not allow for script deviation when administering the dynamic assessment (the examiner simply read exactly what was written on the protocol). Further, the scoring procedures required a simple calculation of gain scores from pretest to posttest, an examiner judgment of modifiability, and a post hoc analysis of the residuum gain score and of the reading strategies applied during the posttest. Because of the highly scripted administration procedures, and relatively simple scoring procedures, we have confidence that the administration fidelity and interrater reliability estimates derived from the follow-up study are transferable to the current data.

One additional consideration is that we do not know the extent to which test-retest effects contributed to the gain scores from pretest to posttest. Although in most testing contexts test-retest effects are undesired, in the context of dynamic assessment, such effects are expected and can be an indicator of learning potential. Nevertheless, future research should explore the degree to which the teaching procedures of the dynamic assessment uniquely contribute to classification accuracy.

**Summary and Conclusions**

Results suggest that the dynamic assessment of reading used in this study may be a useful measure for early identification of bilingual, Latino children who will be at risk for reading decoding difficulties during first grade. This is important because the dynamic assessment is relatively simple to administer, takes less than 5 minutes to complete, and likely includes procedures and content that are more

Table 6. Participant Characteristics of the Attrition Group and the Longitudinal Group and First-Grade Reading Performance for the Longitudinal Group.

<table>
<thead>
<tr>
<th>Measure</th>
<th>n</th>
<th>M</th>
<th>SD</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Longitudinal group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>48</td>
<td>22 (46%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>48</td>
<td>26 (54%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>48</td>
<td>65.1</td>
<td>3.8</td>
<td>64.5</td>
</tr>
<tr>
<td>Socioeconomic status (weighted Hollingshead score)</td>
<td>48</td>
<td>14.6</td>
<td>7.4</td>
<td>14.8</td>
</tr>
<tr>
<td>Years in preschool</td>
<td>48</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>Kindergarten dynamic assessment</td>
<td>48</td>
<td>4.3</td>
<td>1.9</td>
<td>5.0</td>
</tr>
<tr>
<td>First-grade ORF</td>
<td>48</td>
<td>22.9</td>
<td>21.9</td>
<td>16.0</td>
</tr>
<tr>
<td>First-grade NWF</td>
<td>48</td>
<td>58.0</td>
<td>29.7</td>
<td>51.0</td>
</tr>
<tr>
<td>First-grade WRMT-R</td>
<td>37</td>
<td>39.0</td>
<td>13.2</td>
<td>41.0</td>
</tr>
<tr>
<td><strong>Attrition group</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Male</td>
<td>15</td>
<td>7 (47%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>15</td>
<td>8 (53%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (months)</td>
<td>15</td>
<td>66.2</td>
<td>5.3</td>
<td>66.0</td>
</tr>
<tr>
<td>Socioeconomic status (weighted Hollingshead score)</td>
<td>15</td>
<td>13.7</td>
<td>6.7</td>
<td>12.5</td>
</tr>
<tr>
<td>Years in preschool</td>
<td>15</td>
<td>0.8</td>
<td>1.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Kindergarten dynamic assessment</td>
<td>15</td>
<td>4.0</td>
<td>2.0</td>
<td>5.0</td>
</tr>
<tr>
<td>First-grade ORF</td>
<td>0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>First-grade NWF</td>
<td>0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
<tr>
<td>First-grade WRMT-R</td>
<td>0</td>
<td>n/a</td>
<td>n/a</td>
<td>n/a</td>
</tr>
</tbody>
</table>

*Note.* The attrition group is comprised of the students who left the school district prior to the end of first grade testing. The longitudinal group is comprised of those students who were assessed at kindergarten and then again at the end of first grade. ORF = DIBELS first-grade oral reading fluency; NWF = DIBELS first-grade nonsense word fluency; WRMT-R = Woodcock Reading Mastery Test–Revised; DIBELS = Dynamic Indicators of Basic Literacy Skills.
sensitive to a culturally and linguistically diverse population than static measures.

By using the narrow view of reading as the theoretical construct of the formative criterion measures, the results of the dynamic assessment decoding strategy measure are easily interpretable and could lead to highly specific action for those children identified at risk. In conjunction with the classification accuracy, these findings provide early evidence of inferentially derived construct validity and predictive validity of the dynamic assessment that was used in this study.

Appendix A

Dynamic Assessment Modifiability Scoring Guide.

<table>
<thead>
<tr>
<th>Dynamic assessment decoding strategy score (calculated for posttest)</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 point</td>
<td>Whole word guess + no evidence of onset-rime strategy + no evidence of initial sound learning</td>
</tr>
<tr>
<td>2 points</td>
<td>Whole word guess + no evidence of onset-rime strategy</td>
</tr>
<tr>
<td>3 points</td>
<td>Whole word guess + evidence of onset-rime strategy</td>
</tr>
<tr>
<td>4 points</td>
<td>No whole word guess</td>
</tr>
</tbody>
</table>

Response to instruction score (calculated immediately following teaching phase)

Rate how difficult it was for the child to respond to the instruction during the teaching phase:

<table>
<thead>
<tr>
<th>0 points</th>
<th>Difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 point</td>
<td>Moderate</td>
</tr>
<tr>
<td>2 points</td>
<td>Easy</td>
</tr>
</tbody>
</table>

Note. The Dynamic Assessment Modifiability Score was obtained by calculating the sum of the Decoding Strategy Score and the Response to Instruction Score. A Dynamic Assessment Modifiability Score of 2 or lower was used to classify children as at risk for future reading difficulty. If the teaching phase and the posttest were not administered (a child correctly read 75% or more of the sounds or words at pretest), then that child received a Dynamic Assessment Modifiability Score of 6 (the highest score possible).

Bilingualism, language proficiency, and learning to read in two writing systems. Journal of Educational Psychology, 97(4), 580–590.


