Center for
Early Care and Education Research
Dual Language Learners

RESEARCH BRIEF #9

Assessment Tools for Language and Literacy Development of Young Dual Language Learners (DLLs)

IN THE PAST DECADE, a dramatic increase in investments in early childhood programs (Barnett et al., 2011) has been accompanied by an increase in the number of dual language learners (DLLs)—children from birth to age five years who are growing up in homes where languages other than or in addition to English are spoken (Aud et al., 2012; Aikens et al., 2011; Vogel et al., 2011). For example, recent reports from the Family and Child Experiences Survey (FACES) 2009 cohort, the ongoing national study of children in Head Start, indicate that more than 31 percent of preschoolers in the program live in homes where a language other than English is spoken (Aikens et al., 2011). This is also true of almost one-third of children in Early Head Start nationally (Vogel et al., 2011). Reliable and valid assessments for this population are needed for evaluations of early childhood programs and how well they are meeting the needs of DLLs.

All children need assessments that are fair, equitable, and well constructed. The results should represent skills, knowledge, and behaviors of DLLs that are important for their development and later school success. Since the questions and probes for most assessments require the use of a single language, either by the assessor or the child, the results from an assessment of a child who is learning two languages may not represent fully his or her knowledge and skills.

Examination of reliability and validity evidence helps us select measures by providing information about when and with whom the assessment will provide trustworthy and meaningful results. Information about reliability can tell us how dependable the results are, while evidence of validity indicates whether the assessment is measuring what it is supposed to be measuring and what types of inferences can be drawn from the results. Both reliability and validity inform what types of inferences can be drawn from the results.

Typically, information about evidence of the reliability and validity of particular assessments is found in the technical manuals that accompany them (see Table 1 for examples). Other than a few designed specifically for DLLs, most assessments report about the reliability and validity evidence based on samples that are more representative of children who are monolingual and make up the majority of the standardization samples. Test development is expensive, and publishers usually do not invest in seeking evidence of validity for subgroups such as DLLs.

Assessments are validated by accumulating evidence in relation to different types of inferences. An assessment may be valid for a particular purpose or representative of the skills of one group of children but not another. The key issue is whether an assessment really measures what it declares to measure when used with a group of children similar to those in the study and for that particular purpose (for example, program evaluation, research about children's development, or identification of a child's need for more intensive instruction or specialized intervention).

Different types of evidence can be collected to support interpretation of an assessment and the results obtained from it. The normative group (the sample used to develop comparison scores on the assessment) is a key component. An assessment may be valid for comparing children from the same linguistic group but unfair when comparing the skills of children across linguistic groups.

A measure may include items that function in differen ways across linguistic groups, including differences in difficulty, discrimination, and/or factor loadings. The abilities of one group may not be validly represented by the scores derived from the items, redereing cross-group comparisons unfair. In a similar way, an assessment may not be valid for making critical decisions about children or the programs serving those children. For example, an English vocabulary or language assessment would be a valid indicator of whether a child is learning English, but using standard scores based on children who speak only English to determine if a DLL has a language disability would not.

Table 1. Common Types of Reliability and Validity Evidence

Term	Definition
Internal consistency reliability	Indicates whether the items within a measure are measuring the same underlying concept
Test–retest reliability	Indicates whether assessment of a particular construct (what is being measured) would result in the same score if repeated later
Differential item functioning	Notes whether the items within a measure function in the same way for different groups of children
Validity	Indicates whether an assessment is measuring what it purports to measure and under what conditions; assessments can be valid for some uses and not for others
Predictive validity	Indicates whether the assessment is related in expected ways to similar or related outcomes measured at a later time

To extend the available information about assessments used with DLLs, we examined research studies published in the last 10 years that included young DLLs. This brief draws from a more detailed report examining language and literacy measures used in seven large-scale government studies and thirty research studies conducted in the past decade (Bandel et al., 2012).[1] This brief highlights key issues noted in the report and emphasizes questions to consider when evaluating the appropriateness of assessment tools used with DLLs and interpreting research that involves such assessments.

How reliable is the assessment?

Reliability information indicates to consumers of research the trustworthiness of the findings and results can differ by sample and study. Poor reliability can limit the ability to detect change over time or associations between constructs (for example, between vocabulary and early literacy)—particularly when the sample size is small—and can lead to flawed conclusions. Many studies in our review did not provide any study-specific evidence of reliability but reported only the published evidence from the assessment manual. Our review suggests that more information is needed about the performance of measures with DLLs.

Internal Consistency

The most commonly reported indicator of reliability is internal consistency—that is, how consistently the items within an assessment measure the construct of interest. However, most of the peer-reviewed studies did not include study-specific estimates (reporting only the published reliability estimates, if at all) or separate estimates by language group. Large-scale studies were more likely to include sample-specific estimates and had sufficient sample sizes to report estimates by subgroups. When reported, the estimates of internal consistency in the studies reviewed were generally favorable, though not always within acceptable ranges for Spanish assessments. For example, the reported reliability estimates for the Spanish version of the Story and Print Concepts used in the FACES 2000 and FACES 2006 studies were much lower than the reliability estimates for the English version.

Different factors can affect the reliability of measures including how targeted the items are to the construct being measured and to the sample of children taking the assessment, how well items discriminate among children, and the number of items. Usually, measures with more items will demonstrate greater reliability, though well-constructed adaptive assessments can attain highly

reliable estimates with a limited number of items. Many of the assessments were adaptive assessments with items targeted to the children's developmental levels. However, when compared with an English dominant sample, there may be greater differences in experiences, languages, and dialects among DLLs that affect their responses to items in ways unrelated to what is being measured. In those cases, more items may be needed to obtain reliable estimates. Though often weaker for Spanish assessments, reported internal consistency estimates were usually within acceptable ranges when a greater number of items was administered. The weaker estimates of reliability on some Spanish assessments relative to the English assessments may be due in part to the number of items administered. For example, in FACES 2009, the reliability estimates for the Spanish literacy assessments in the fall were weak (less than .70), but they were based on an average of only 14 administered items, compared to the average of 17–26 items administered to the children (including some DLLs) taking the assessments in English, which had stronger internal consistency estimates (about .80). Because many assessments used in early childhood research are adaptive measures—that is, they present items of increasing difficulty until the child incorrectly responds to a specified number of items—the number of items administered differs according to the child's responses and the stop rules of the assessment (this is elaborated on below). More research is needed specifically about the validity of stop rules themselves for differing populations—in particular, how reliable and valid they are when administering assessments to DLLs.

Stability

The stability of a score, or test–retest reliability, indicates whether an assessment of a particular construct would result in the same score if repeated a week later or, for more stable constructs, months or years later. Among the studies we reviewed, test–retest reliability within a four-week time period was reported for only one bilingual measure, the *Bilingual English Spanish Oral Language Screener* (BESOS; Peña, Bedore, Gutierrez-Clellen, Iglesias, & Goldstein, in preparation), with slightly stronger reliability indicated for the Spanish than for the English version.

Evidence of stability across longer periods of time (more than three months) was available for several measures of vocabulary and literacy in the reviewed studies, but these may not be a good measure of stability particularly during the preschool years when DLLs are likely to have increased exposure to English via educational opportunities. For some children, preschool may be the first educational experience and the first exposure for some DLLs to English. This makes it very difficult to disentangle differentiate the stability of the assessment from the responsiveness of different children to the educational opportunities and the sensitivity of the measure to intervention. Children who are learning English for the first time might be expected to show greater change in their knowledge of English words than children who have been learning it for an extended period of time. Thus, the stability estimates over a long period of time would be lower than over a two-week period, particularly when samples have both simultaneous and sequential DLLs. With that caveat in mind, examination of correlations suggested that literacy measures might be less stable than vocabulary measures (Dickinson, McCabe, Clark-Chiarelli, & Wolf, 2004; Hammer, Davison, Lawrence, & Miccio, 2009), although this was not always the case (Anthony et al., 2009). However, the samples and the length of time between assessments varied, making it more difficult to draw valid inferences from the estimates.

How fair is the assessment?

Bias from a number of sources may render an assessment unfair or invalid for a particular group of children. The questions or probes may use formats or words that are unfamiliar in some cultures or the questions may sample knowledge that is specific to a certain culture or linguistic group. Under these conditions, the items would not be representative of the experiences and knowledge of some of the children assessed. In short, the evidence suggests that some approaches to assessing young DLLs may not result in a fair representation of the children's knowledge. For example, assessing the conceptual vocabulary of a DLL—that is, whether a child has words for different objects, actions, and concepts—in a single language is likely to under-represent the

words the child knows. Children typically acquire words for objects and activities experienced at home in the language used most often at home and words for academic concepts in the language used in school (Bialystok et al. 2010). Using one of the most commonly applied measures of English vocabulary (PPVT-III), Bialystok and colleagues (2010) noted that test items referring to home objects and activities were more difficult for Spanish-dominant children than for English-dominant children, while school-related words posed similar difficulty across groups.

Basal and Ceiling Rules

Implications of this phenomenon may go beyond the correctness of individual items when assessments of young children are adaptive, using basal and ceiling (or start and stop) rules. The items in adaptive assessments, as mentioned earlier, are generally ordered in terms of difficulty based on the responses of the normative sample, and the stop rules are designed so children will have a very low probability of getting any items beyond the stopping point correct; the scoring assumes responses to all subsequent items are incorrect. Because items concerning the names of objects found in the home are easier for young children to correctly name, they are presented early in assessments of vocabulary. Children who are DLLs, however, may know the English names for objects and activities related to school and academics, but not for those associated with the home. Using the published ceiling rules could, therefore, result in underestimates of their English vocabulary. None of the reviewed studies examined the appropriateness of basal and ceiling rules for young DLLs.

Differential Item Difficulty

Greater attention is needed for evaluating the fairness of individual items and tasks across subgroups. Some tasks may not have the same meaning in development across subgroups or the difficulty may vary across languages. For example, rhyming is usually easier for young children when the words have only one syllable, but English has many more one-syllable rhymes than Spanish. Limited evidence was provided in the reviewed studies for the congruence of estimates of item- difficulty across languages, though many of the reviewed research studies included children who spoke primarily English and children who spoke primarily Spanish or another language. Although 74 percent of the studies assessed children in both languages, only one (Bialystok, Luk, Peets, & Yang, 2010) reported examination of how items functioned for different groups. Particularly when researchers translate items or develop their own assessments, attention is needed to assure that the assessments are fair representations of children's skills, knowledge, and ability. Test developers and researchers should evaluate the items in tests to determine if they have the same meaning across groups—that is, indicating whether the items function in the same way for different groups of children.

Parent and Teacher Reports

Some studies, particularly for children younger than two years, used assessments based on parent or teacher reports. Early childhood researchers often rely more on teacher reports than parent reports, due to efficiencies in data collection and potential systematic bias with parent reporting (Pan Rowe, Spier, Tamis-LeMonda, 2004; Rescorla, Ratner, & Jusczyk, & Jusczyk, 2005; Reese & Read, 2000; Roberts Burchinal, & Durham, 1999). With English monolingual samples, correlations of teacher report with direct assessments range from moderate to high during the elementary school years, but suggest potential biases with teacher reports particularly for items that have higher inference levels (Perry & Meisels, 1996). Less research is available for the infant-toddler years and the research with DLLs is particularly limited in this regard. However, a study by Vagh and colleagues (Vagh, Pan, & Mancilla-Martinez, 2009) suggests that parents (rather than teachers) may provide the most valid information about DLLs' vocabulary development. Among DLLs, stronger relationships were found between parent reports of children's vocabulary and direct assessment of vocabulary than between teacher reports of those children's vocabulary and direct assessment of their skills (Vagh, Pan, & Mancilla-Martinez, 2009).

Conceptual Scoring

Other studies used measures that are conceptually scored. Conceptual scoring allows probes or responses in more than one language and will usually present

a more valid assessment when measuring children's knowledge of concepts rather than vocabulary in a particular language. For receptive vocabulary, however, comparisons across linguistic groups can be more challenging. If you present four pictures of objects and the name of one of them in one language (giving a 1 in 4 chance of selecting the correct picture) and the child is unsuccessful, then follow with a prompt in the other language (now giving a 1 in 3 chance of success), the probability of success favors children who know two languages over children who know only one. This is not a problem for expressive vocabulary because a child could draw upon thousands of different words to name a picture.

It is important that researchers consider all sources of potential bias when presenting and interpreting their findings; and, when feasible, contribute to the knowledge base about validity of measures with DLLs.

Does the Assessment Answer the Questions of Interest?

If assessments of DLLs are to be valid for the goals of the research, care must be taken to select instruments and methods that match the question to be answered. For instance, it is important to consider whether the research question requires measurement of skills, knowledge, or behavior at a single point in time or if the focus is on change, particularly if the change may reflect the effectiveness of a program or intervention. In the latter case, researchers should consider the language used in the program or intervention and the possibility that they will need to assess the constructs of interest at baseline in both the child's home language and the language used in the program to have a valid estimate of the program's effects. Without an assessment in the home language, children may appear to lack skills at baseline and show gains in many skills when assessed at a later point in time, when the only knowledge that they acquired was comprehension of the questions in English. A program with many DLLs could appear to be making greater gains than a program in which children enter without any skills at all. For example, if the children already possess understanding of a concept such as seriation (for example, tall, taller, and tallest), but cannot understand the question, they would do poorly on those items at baseline, but do well once they understand the question in English. The change was in their understanding of English rather than their understanding of seriation. In another program, children may enter without understanding of seriation and so would need to learn both the English and seriation. At the end of the year, when children in both groups successfully responded to the seriation tasks, the program that taught the children the English, but did not extend mathematical thinking, would look as though it had a greater effect on cognitive development than it actually did.

What Levels of Performance and Progress Are Expected for DLLs on Different Assessments?

More information is needed about expected performance of DLLs on assessments. To provide information about their expected performance on language measures, developers should provide supplemental norms for DLLs or estimates of the mean and standard deviation for the subsample of DLLs. Many samples of DLLs in the studies we reviewed were relatively small, and often the only norms the researchers had available for interpreting the scores were based on standardization samples in which most of the children were monolingual in either English or Spanish.

Most research studies conducted with DLLs in the United States in the last 10 years included primarily or only low-income, Spanish–English DLLs, limiting the generalizability of the findings to other groups of DLLs. Nationally, however, the majority of young DLLs reside in homes with limited income and have Spanish as a home language (Shin and Kominski, 2010).

It would be helpful if researchers (particularly for large-scale studies) provided separate information about the mean scores and study characteristics of their DLL samples (including socioeconomic background, range of dialects represented, and age range) and their skills.

Vocabulary Assessments Used with DLLs in Reviewed Studies

Assessment Abrieviations	Title, Author, Date
English Vocabulary	
CDI	MacArthur-Bates Communicative Development Inventories (Fenson et al., 1993)
EOWPVT	Expressive One-Word Picture Vocabulary Test (Brownell, 2000a)
ROWPVT	Receptive One-Word Picture Vocabulary Test (Brownell, 2000b)
PPVT	
PPVT–R	Peabody Picture Vocabulary Test–Revised (Dunn & Dunn, 1981)
PPVT–III	Peabody Picture Vocabulary Test–III (Dunn & Dunn, 1997)
PPVT–4	Peabody Picture Vocabulary Test–4 (Dunn & Dunn, 2007
WJ–III (Picture Vocabulary)	Woodcock Language Proficiency Battery–Revised, English Form (Woodcock, 1995)
Spanish Vocabulary	
Inventario (CDI Spaish Edition)	El Inventario del Desarrollo de Habilidades Comunicativas (Jackson-Maldonado, Thal, Marchman, Newton, Fenson, & Conboy, 2003)
TVIP	Test de Vocabulario en Imágenes Peabody (Dunn, Padilla, Lugo, & Dunn, 1986)
WM–II (Vocabulario Sobre Dibujos)	Woodcock Language Proficiency Battery–Revised, Spanish Form (Woodcock & Muñoz-Sandoval, 1995)
Conceptuall Scored (English & Spanish)	
EOWPVT-SBE	Expressive One-Word Picture Vocabulary Test: Spanish Bilingual Edition (Brownell, 2001a)
ROWPVT-SBE	Receptive One-Word Picture Vocabulary Test : Spanish Bilingual Edition (Brownell, 2001b)
SEVC	Spanish–English Vocabulary Checklist (Patterson, 1998)
Chiese Vocabulary	
PPVT–R Chinese	Peabody Picture Vocabulary Test—Revised: Chinese (Lu & Liu, 1998)

What Does the Score Mean and How Can We Interpret the Results?

Study results should be interpreted and discussed with reference to the characteristics of the children in the study. When comparing differences in the performance and progress of groups of children, information about differences in opportunity to learn (due to different socioeconomic, cultural, or linguistic backgrounds) should be included. When discussing standard scores, researchers should help readers understand the similarities and differences between the normative group and the study sample. Even for the same sample of children, mean standard scores vary across different measures of the same construct.

The baseline information in FACES 2009 (Aikens et al., 2011), as one example, provides potentially different interpretations of DLLs' knowledge of vocabulary and concepts depending on the dimensions assessed and the norm group used for deriving standard scores. FACES 2009 assessed vocabulary—the area most commonly assessed among young DLLs (see Table 2)—using multiple measures and provided standard scores for 4-year-old DLLs based on the PPVT-4 (receptive English vocabulary), the English version of the EOWPVT (expressive vocabulary), and the Spanish bilingual version (EOWPVT-SBE), in addition to scores on the TVIP, a Spanish receptive vocabulary measure. Although mean scores

for the predominantly English-speaking DLLs were in a similar range across the English assessments, the scores for the Spanish-speaking DLLs varied more widely on the English and bilingual assessments, with mean standard scores of 56, 67, and 86 on the PPVT-4, and the English and bilingual norms for the EOWPVT-SBE, respectively. Standard scores for Spanish- speaking DLLs were more similar between the TVIP and the EOWPVT-SBE (81 and 86, respectively). These latter two measures have standard scores based on Spanish-speaking or bilingual samples of children, with half or more of the sample from homes with limited maternal education. These differences and similarities highlight the difficulty in interpreting children's skills without information about the assessments and the normative samples used to generate the standard scores.

Do the Assessments Measure What They Are Supposed to Measure?

The studies provided some additional evidence of validity—that is, that the assessments measured what they were supposed to measure. The type of evidence most frequently identified for language and literacy assessments of DLLs was an association with children's age or exposure to English, including moderate correlations of vocabulary and language assessments with age and parent-reported exposure to English, as well as evidence of an increase in assessment scores across time.

Similar to reliability coefficients, the range of reported concurrent validity coefficients was weaker when compared to coefficients found across measures in studies of young monolingual English samples.[4] They ranged from low-moderate to strong, with stronger estimates between vocabulary and language measures and weaker estimates between vocabulary and literacy measures.

Do the Assessments Support Understanding of the School Readiness of DLLs?

Very limited evidence was available for the predictive validity of early measures for later outcomes when used with DLLs. Most evidence of validity of the measures used in the reviewed studies was found with samples of children who were able to take assessments in English, and in those studies, the DLLs were combined with English-only speakers. To better understand how to support DLLs and monitor progress toward school readiness, separate validity analyses with DLLs are needed, as well as more information about how the Spanish versions of assessments inform our understanding of how children are progressing toward success in school.

None of the reviewed studies provided any estimates of concurrent or predictive validity separately for monolingual and DLL children. The sample sizes of the peer-reviewed research studies were often too small for separate subgroup analyses of validity. Other than the large-scale government studies, only three of the research studies had sample sizes greater than 200. Predictive validity evidence in these studies usually examined whether children's scores increased across time points and whether earlier vocabulary and language assessments predicted later literacy.

The studies that provided information about the predictive validity of the measures in relation to school readiness did not examine findings for measures separately by subgroup. The report for FACES 1997 (Zill et al., 2003) only included children who took the English assessment at each time point and did not clearly indicate how many of them had Spanish as a home language. One study of DLLs reported analysis of predictive relationships from preschool to first grade reading in English with the full sample and did not estimate differences by language proficiency (Rinaldi & Páez, 2008). With the full sample, that study indicated weak relationships for individual subtests, although a combination of several subtests across both Spanish and English increased the amount of explained variance in English reading. Hammer and colleagues (2007) used a combination of language measures to examine the relationship between preschool language development and spring kindergarten reading, comparing the skills of DLLs who learned English at home prior to starting preschool (Head Start) to those of DLLs who were introduced to English at Head Start. Although the analyses provided estimates for each of the subgroups, the evidence cannot be attributed to a single measure, but rather to the component measures.[5]

Many factors can affect the strength of validity coefficients in early childhood. Typically, the strength of the

relationship between assessments will depend on many different factors, including the reliability of the assessment, similarity in mode of assessment, similarity in the dimensions and constructs assessed, time between assessments, and age of the child (the younger the child at first assessment, the weaker the relationship). The predictive validity evidence of early childhood assessments among English monolingual samples typically demonstrates low to moderate predictive correlation coefficients, with less than 25 percent of the overall variance in early academic performance predicted from any single preschool measure (LaParo & Pianta, 2000). When samples include DLLs, the number of additional variables that can affect the strength of the coefficient increases—for example, the age of introduction to the language used, the amount of exposure to the language of assessment, and intervention or preschool experiences. Of the studies reviewed, vocabulary was assessed most often, yet evidence of a relationship to school outcomes was found only with latent traits combining multiple aspects of language (Hammer, Lawrence, and Miccio, 2007; Rinaldi and Páez, 2008). More research is needed to determine how to monitor the development of DLLs to ensure later success in school.

Conclusions

More information is needed about our current assessments and what inferences can be drawn from their results. Also needed are assessments for use with DLLs who speak languages other than English. Researchers should consider carefully what assessment will answer the questions of interest and report more information about the samples and measures.

- Researchers need to consider if assessments are valid for the children in their samples. When samples include children from multiple linguistic backgrounds, will the methods and items fairly represent all children's knowledge, skills, and behaviors? Are cultural or linguistic biases inherent in the use of the assessment with a particular group of children? Even within a single linguistic group, differences in dialect may bias results unless the assessment accounts for them.
- Researchers should consider whether the selected assessment(s) are biased in any way: Do the tasks or items require similar levels of skill across languages and cultures? Is the task equally representative of skills across different groups?
- Test developers (including researchers who develop forms in other languages) should provide evidence that the items are equivalent for children from different groups, that is, that the items contribute to measurement of the construct in the same ways across groups and that the difficulty of each of the items or tasks is the similar across groups.
- More predictive validity evidence is needed, particularly for infant-toddler measures. While this is true for early childhood assessments in general, measures and evidence of predictive validity are particularly sparse for young DLLs.

(Endnotes)

1. The review included peer-reviewed journal articles published between 2000 and 2010 with samples from the United States or its territories and Canada that included at least one direct child assessment or standardized rating of the language or literacy development of DLL children prior to age 6 or kindergarten entry. To examine psychometric properties, we excluded studies with less than 25 DLLs and those that only analyzed language samples, resulting in 30 peer-reviewed research articles. Spanish and English were the most commonly reported languages in the samples, and more studies focused on preschoolers than on infants and toddlers. We also reviewed government reports of large-scale studies of early childhood that included at least one direct child assessment of language and/or literacy. Among government reports published in the last 10 years, we located only seven large-scale national studies that examined children's language or literacy development prior to kindergarten entry and included DLLs in any of these assessments. Description of the sample characteristics and study purpose for each of the studies can be found in the full report (Bandel et al., 2012) which is available on the CECER-DLL website [http://cecerdll.fpg.unc.edu].

2. Researchers and assessment developers often require that assessment tools have evidence of reliability values of 0.70 or higher to support inferences about the measure (Bacon 2004; Cohen 1977; Litwin 2003; Nunnally 1978), however, the minimal recommended level of internal consistency differs according to the type of inference that will be made about the results.

3. The reported reliability estimates for the Spanish version of the Story and Print Concepts used in the FACES 2000 and FACES 2006 studies were less than .60, while reliability estimates for the English version were greater than .70.

4. Correlations with the DLL samples ranged from .25 to .79. Among English-only samples, estimates are often stronger. Some examples of correlations among English assessments include the PLS-5 and the CELF P-2, with r = .79 for total scores; r = .82 for expressive language scores from each of these assessments. Scores on the PPVT-4 (receptive vocabulary) with the Expressive Vocabulary Test-Second Edition (EVT-2) ranged from .80 to .84.

5. The English language component measure was based on the PPVT-III and the Receptive Language subtest of the TELD-3. The Spanish language component measure was based on the TVIP and the Auditory Comprehension subtest of the PLS-3. The study found that the growth on the component measure during the preschool year predicted literacy scores at the end of kindergarten.

References

Aikens, N., Hulsey, L. K., Moiduddin, E., Kopack, A., Takyi-Laryea, A., Tarullo, L., & West, J. (2011). Data tables for FACES 2009 Head Start children, families, and programs: Present and past data from FACES report (ACF-OPRE 2011-33b). U.S. Department of Health and Human Services, Office of Planning, Research, and Evaluation, Administration for Children and Families. Washington, DC: U.S. Government Printing Office.

Anthony, J. L., Solari, E. J., Williams, J. M., Schoger, K. D., Zhang, Z., Branum-Martin, L., & Francis, D. J. (2009). Development of bilingual phonological awareness in Spanish-speaking English language learners: The roles of vocabulary, letter knowledge, and prior phonological awareness. *Scientific Studies of Reading, 13*(6), 535–564.

Aud, S., Hussar, W., Johnson, F., Kena, G., Roth, E., Manning, E., … Zhang, J. (2012). *The condition of education 2012* (NCES 2012-045). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Bacon, D. (2004). The contributions of reliability and pretests to effective assessment. *Practical Assessment, Research & Evaluation, 9*(3).

Bandel, E., Atkins-Burnett, S., Castro, D. C., Wulsin, C. S., & Putnam, M. (2012). *Examining the use of language and literacy assessments with young dual language learners*. Research report no. 1. Center for Early Care and Education Research–Dual Language Learners (CECER-DLL). Chapel Hill: University of North Carolina, Frank Porter Graham Child Development Institute.

Barnett, W. S., Carolan, M. E., Fitzgerald, J., & Squires, J. H. (2011). *The state of preschool 2011: State preschool yearbook*. New Brunswick, NJ: National Institute for Early Education Research.

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism: Language and Cognition, 13*(4), 525–531.

Brownell, R. (2000a). *Expressive one-word picture vocabulary test*. Novato, CA: Academic Therapy Publications.

Brownell, R. (2000b). *Receptive one-word picture vocabulary test*. Novato, CA: Academic Therapy Publications.

Brownell, R. (2001a). *Expressive one-word picture vocabulary test* (Spanish-Bilingual Edition, Manual). Novato, CA: Academic Therapy Publications.

Brownell, R. (2001b). *Receptive one-word picture vocabulary test* (Spanish-Bilingual Edition, Manual). Novato, CA: Academic Therapy Publications.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic Press, 1977.

Dickinson, D. K., McCabe, A., Clark–Chiarelli, N., & Wolf, A. (2004). Cross-language transfer of phonological awareness in low-income Spanish and English bilingual preschool children. *Applied Psycholinguistics, 25*(03), 323–347.

Dunn, L., & Dunn, L. (1981). *Peabody picture vocabulary test—revised*. Circle Pines, MN: American Guidance Service.

Dunn, L., & Dunn, L. (1997). *Peabody picture vocabulary test* (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, D., & Dunn, L. (2007). *Peabody picture vocabulary test* (4th ed.). Minneapolis, MN: NCS Pearson, Inc.

Dunn, L., Padilla, E. R., Lugo, D. E., & Dunn, L. (1986). *Test de vocabulario en imagenes Peabody*. Circle Pines, MN: American Guidance Service.

Fenson, L., Dale, P., Reznick, S., Thal, D., Bates, E., Hartung, J., … Reilly, J. (1993). *MacArthur communicative development inventories*. San Diego, CA: Singular Publishing.

Hammer, C. S., Lawrence, F. R., & Miccio, A. W. (2007). Bilingual children's language abilities and early reading outcomes in Head Start and kindergarten. *Language, Speech, and Hearing Services in Schools, 38*(3), 237.

Hammer, C. S., Davison, M. D., Lawrence, F. R., & Miccio, A. W. (2009). The effect of maternal language on bilingual children's vocabulary and emergent literacy development during Head Start and kindergarten. *Scientific Studies of Reading, 13*(2), 99–121.

Jackson-Maldonado, D., Thal, D. J., Marchman, V., Newton, T., Fenson, L., & Conboy, B. (2003). *El inventario del desarrollo de habilidades comunicativas: User's guide and technical manual*. Baltimore, MD: Brookes Publishing Co.

LaParo, K. M., & Pianta, R. C. (2000). Predicting children's competence in the early school years: A meta-analytic review. *Review of Educational Research, 70*(4), 443–484.

Litwin, M. S. (2003). *How to assess and interpret survey psychometrics* (2nd ed.). Thousand Oaks, CA: Sage Publications, 2003.

Lu, L., & Liu, H. S. (1998). *The Peabody picture vocabulary test—revised* (Chinese). Taipei, Taiwan: Psychological Publishing.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Pan, B. A., Rowe, M. L., Spier, E., & Tamis-Lemonda, C. (2004). Measuring productive vocabulary of toddlers in low-income families: concurrent and predictive validity of three sources of data. *Journal of Child Language, 31,* 587-608.

Patterson, J. (1998). Expressive vocabulary development and word combinations of Spanish–English bilingual toddlers. *American Journal of Speech–Language Pathology, 7*(4), 46–56.

Peña, E., Bedore, L., Gutierrez-Clellen, V., Iglesias, A., & Goldstein, B. (in preparation). *Bilingual English–Spanish assessment*.

Perry, N. E., & Meisels, S. J. (1996). *How accurate are teacher judgments of students' academic performance?*(NCES 9608). Washington, DC: National Center for Education Statistics.

Rinaldi, C., & Páez, M. (2008). Preschool matters: Predicting reading difficulties for Spanish-speaking bilingual students in first grade. *Learning Disabilities, 6*(1), 71.

Rescorla, L., Ratner, N. B., Jusczyk, P., & Jusczyk, A. M. (2005). Concurrent validity of the language development survey: associations with the MacArthur-Bates communicative development inventories: Words and sentences. *American Journal of Speech and Language Pathology, 14*(2), 156-163.

Reese, E., & Read, S. (2000). Predictive validity of the New Zealand MacArthur Communicative Development Inventorty: Words and Sentences. *Journal of Child Language, 27*(2), 255-266.

Roberts, J. E., Burchinal, M., & Durham, M. (1999). Parents' report of vocabulary and grammatical development of African American preschoolers: Child and environmental associations. *Child Development 70,* 92–106.

Shin, H. B., & Kominski, R. A. (2010). *Language use in the United States: 2007*. American Community Survey Reports (ACS-12). U.S. Census Bureau. Washington, DC: U.S. Government Printing Office.

Vagh, S. B., Pan, B. A., & Mancilla-Martinez, J. (2009). Measuring growth in bilingual and monolingual children's english productive vocabulary development: The utility of combining parent and teacher report. *Child Development, 80*(5), 1545-1563.

Vogel, C. A., Boller, K., Xue, Y., Blair, R., Aikens, N., Burwick, A., & Stein, J. (2011). *Learning as we go: A first snapshot of Early Head Start programs, staff, families, and children* (ACF-OPRE 2011-7). U.S. Department of Health and Human Services, Office of Planning, Research, and Evaluation, Administration for Children and Families. Washington, DC: U.S. Government Printing Office.

Woodcock, R. W. (1995). *Woodcock language proficiency battery—revised*. Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Muñoz-Sandoval, A. F. (1995). *Woodcock language proficiency battery—revised—Spanish form*. Itasca, IL: Riverside Publishing.

Zill, N., Resnick, G., Kim, K., O'Donnell, K., Sorongon, A., Hubbell McKey, R., with others (2003). *Head Start FACES 2000: A whole-child perspective on program performance*. U.S. Department of Health and Human Services, Office of Planning, Research, and Evaluation, Administration for Children and Families. Washington, DC: U.S. Government Printing Office.

About CECER-DLL

CECER-DLL is a national center that is building capacity for research with dual language learners (DLLs) ages birth through five years. CECER-DLL aims to improve the state of knowledge and measurement in early childhood research on DLLs, identify and advance research on best practices for early care and education programming, and develop and disseminate products to improve research on DLLs. CECER-DLL is a cooperative agreement between the Frank Porter Graham (FPG) Child Development Institute at The University of North Carolina at Chapel Hill and the Office of Planning, Research, & Evaluation (OPRE) in the Administration for Children & Families (ACF), in collaboration with the Office of Head Start and the Office of Child Care.

Suggested citation

Atkins-Burnett, S., Bandel, E., & Aikens, N. (2012). *Research brief #9. Assessment tools for the language and literacy development of young dual language learners (DLLs)*. Chapel Hill: The University of North Carolina, FPG Child Development Institute, CECER-DLL.

This brief summarizes results from a review of the literature sponsored by CECER-DLL conducted by a research team consisting of Eileen Bandel, Sally Atkins-Burnett, Dina C. Castro, Claire Smither Wulsin and Marisa Putnam, with Margaret Burchinal, Lisa Lopéz, Vera Gutiérrez-Clellen, and Ellen Peisner-Feinberg as research partners. The work was supported by a cooperative agreement funded by the Office of Planning, Research, and Evaluation (OPRE), U.S. Department of Health and Human Services. Permission to copy, disseminate, or otherwise use information from this document for educational purposes is granted, provided that appropriate credit is given.

Additional Resources: For additional information regarding this research brief, see http://cecerdll.fpg.unc.edu

UNC
FRANK PORTER GRAHAM
CHILD DEVELOPMENT INSTITUTE