Hammill Institute on Disabilities

Topics in Early Childhood Special Education
31(2) 89–98
© Hammill Institute on Disabilities 2011
Reprints and permission: http://www.sagepub.com/journalsPermissions.nav
DOI: 10.1177/0271121410391108
http://tecse.sagepub.com
SAGE

Validity Study of the *Preschool Language Scale–4* With English-Speaking Hispanic and European American Children in Head Start Programs

Cathy H. Qi[1] and Scott C. Marley[1]

Abstract

The purpose of the study was to examine the psychometric properties of the *Preschool Language Scale–4* (PLS-4) with a sample of English-speaking Hispanic and European American children who attended Head Start programs. Participants were 440 children between the ages of 3 and 5 years (52% male; 86% Hispanic and 14% European American). Participants were administered the PLS-4 and *Peabody Picture Vocabulary Test–III* (PPVT-III). The Auditory Comprehension and Expressive Communication subscales and Total Language scale scores for the PLS-4 in this sample had excellent reliability (Kuder Richardson–20s > .90). Validity evidence for the PLS-4 was present, with both subscales being positively correlated with PPVT-III scores. Agreement analysis between the PLS-4 and the PPVT-III indicated that the PLS-4 was less likely to identify a child as having a potential language delay than was the PPVT-III. The results largely support the validity of the PLS-4 for its intended purpose of assessing language skills with preschoolers.

Keywords

PLS-4, language assessment, reliability and validity, preschool, Head Start

Accurate and early identification of preschool children with language delays from low-income backgrounds has been a challenging issue for several decades. Children from low-income families have been shown to score lower than the general population on standardized language tests (Champion, Hyter, McCabe, & Bland-Stewart, 2003; Qi, Kaiser, Milan, & Hancock, 2006; Qi, Kaiser, Milan, Yzquierdo, & Hancock, 2003; Restrepo et al., 2006; Stanton-Chapman, Chapman, Kaiser, & Hancock, 2004; Washington & Craig, 1999). For example, children in Head Start have been found to score approximately one standard deviation (*SD*) below national norms on receptive vocabulary (The Head Start Family and Child Experiences Survey [FACES], 2008). Researchers looking at the performance of children on standardized tests have found comparable performance between African American and European American children from low-income families (Qi et al., 2003; Qi et al., 2006) and between English-speaking Hispanic and European American children enrolled in Head Start (Qi & Marley, 2009). Young Hispanic children "constitute an urgent demographic imperative" (Garciá & Jensen, 2009, p. 3). English-speaking Hispanic children from low-income families have performed lower on the standardized language tests than the standardization sample (Qi & Marley, 2009). These disparities in performance between children from low-income families and the standardization sample on language tests can potentially result in overrepresentation of children from low-income backgrounds in special education programs.

The clinical literature has shown that standardized tests, within their limitation, provide scores that may classify an individual as having typical language skills or as having language skills typically below the normative range (Hegde & Maul, 2006). These tests are also widely used by speech and language pathologists (SLPs) as one of the primary measures for identification of language disorders (Huang, Hopkins, & Nippold, 1997). Because of the critical role standardized tests play in the process of identification, it is important to examine the reliability and validity evidence supporting the use of these tests as indicators of language delay for young children, particularly those raised in poverty.

Construct validity is described as a property inherent in test scores for the scores' intended usage (American Educational Research Association, American Psychological

[1] University of New Mexico, Albuquerque, NM, USA

Corresponding Author:
Cathy H. Qi, Hokona Hall 257, Department of Educational Specialties, University of New Mexico, Albuquerque, NM 87131, USA
E-mail: hqi@unm.edu

Association, & National Council on Measurement in Education [AERA, APA, & NCME], 1999). It refers to "the underlying theory on which the instrument is based" (Kaderavek, 2011). It is important for researchers and SLPs to understand the logical theory that underlies the construction of the instrument (Kaderavek, 2011). It is critical that construct validity of scores be assessed with diverse populations and for the proposed uses of a measure. Construct validation is an ongoing process that necessitates evaluating the evidence supporting a measure in different populations and contexts. It should be noted that the current conception of validity, and by association, reliability, holds that these values are sample and context dependent (AERA, APA, & NCME, 1999).

The *Preschool Language Scale–4* (PLS-4; Zimmerman, Steiner, & Pond, 2002) is a norm-referenced instrument designed to assess the auditory and expressive language abilities of children from birth to 6 years 11 months. The PLS-4 consists of two subscales, Auditory Comprehension (AC) and Expressive Communication (EC). It is a revised version of the *Preschool Language Scale–3* (PLS-3; Zimmerman, Steiner, & Pond, 1992), which has been used widely by SLPs, special educators, and researchers to identify potential language disorders or language delays. It is also used to measure change in language skills over time (Zimmerman & Castilleja, 2005; Zimmerman et al., 2002). The PLS-4 standardization sample included 1,564 children, from ages 2 days to 6 years 11 months. Sampled by primary caregivers' education level, 17% had 11 or fewer years of education, 32% had 12 years, 28% had 13 to 15 years, and 23% had 16 or more.

The PLS-4 differs from the earlier version, the PLS-3, in several ways. First, the PLS-4 does not use specific cutoff scores to classify mild, moderate, or severe impairment as did the PLS-3. The reason for this lack of specification is to allow consumers to use different cutoff scores based on local norms or experience (Zimmerman & Castilleja, 2005). Second, the PLS-4 was normed with a nationally representative group in 2001. This renorming resulted in greater representation of minority populations. Of the norm group, 39.1% were ethnic/racial minorities on the PLS-4 while the PLS-3 sampled only 31% minorities. The PLS-4 standardization sample included 18.1% Hispanic. Third, the PLS-4 standardization sample included 13.2% participants with "identified conditions/diagnoses" (Zimmerman et al., 2002, p. 179), among which 5.4% had articulation disorders and 1.9% had language disorders, while the PLS-3 had insufficient numbers of children with language delays. Fourth, the scoring for the PLS-4 has been designed to "pass" items that would be considered "correct" in four dialect patterns (i.e., African American English, Southern English, Appalachian English, and English influenced by another language) even if that item would be considered "incorrect" in Mainstream American English. This modification was intended to reduce the overidentification of children with a potential language delay due to dialectic differences. Finally, the PLS-4 used only a subsample of the children with "language disorders" in their study (those with the most severely impaired expressive language), whereas the PLS-3 study sample included the entire "language disordered" sample. Under these conditions, one would expect better sensitivity and better specificity in the PLS-4. Although researchers and clinicians have expressed concern that the PLS-4 is "inflating scores," the PLS-4 manual reported better discriminant validity than did the PLS-3. One explanation might be that the cutoff score for the PLS-3 was 77 (1.5 *SD* below the mean), whereas the cutoff score for the PLS-4 was 85 (1 *SD* below the mean). Thus, children would have to score lower on the PLS-3 to be classified as having "language disorders" than on the PLS-4. This change in cutoff scores alone should result in greater sensitivity (i.e., the correct identification of children with language delays) because a larger percentage of children would score 1 *SD* below the mean and require further evaluation. However, this improvement in sensitivity comes at a cost. Namely, the specificity (i.e., the correct identification of children with typical language development) of the measure will decline and a greater number of children with typical language development will be identified for further evaluation (Personal communication with P. Yoder, October 18, 2004).

When researchers or SLPs determine the accuracy of a diagnostic test, they want to determine the extent to which individuals with typical language development can be distinguished from those with language disorders on the basis of test scores (Maxwell & Satake, 2006). Sensitivity can be defined as the probability that the test indicates an individual has language disorders, when in fact he or she does have language disorders. Specificity can be defined as the probability that the test indicates an individual does not have language disorders, when in fact the language disorders do not exist. The sensitivity of the PLS-3 total scores for 3-, 4-, and 5-year-olds are .36, .61, and .45, respectively, and the specificity of the PLS-3 total scores for 3-, 4-, and 5-year-olds are .96, 1.0, and .97, respectively. However, the sensitivity of the PLS-4 total scores for 3-, 4-, and 5-year-olds are .83, .78, and .79, respectively, and the specificity of the PLS-4 for 3-, 4-, and 5-year-olds are .88, .96, and .82, respectively. The PLS-4 developers may have systematically designed their study to give better sensitivity than the PLS-3 study by selecting a less rigorous cutoff score (Personal communication with P. Yoder, October 18, 2004; Zimmerman et al., 1992; Zimmerman et al., 2002).

In addition to the above changes, evidence for the PLS-4 scoring is generally compelling. First, the PLS-4 has strong evidence of reliability, with high internal consistency, test–retest reliability, and intertester agreement (Zimmerman & Castilleja, 2005). Second, data reported in the Examiner's Manual are supportive of the content validity and concurrent validity. For example, a comparison of PLS-4 scores with the *Denver II* (Frankenburg et al., 1990) and the PLS-3 yielded high intermeasure correlations. The correlations between the PLS-3 and the PLS-4 were .65 for the AC standard scores, and .79 for the EC standard scores. These moderate to high correlations support the contention that the measures are assessing similar constructs.

Despite the positive changes made by the PLS-4, several concerns were raised about the test (Suen, 2004). First, clinical impressions indicated that more children might be identified as having potential language delays through the use of the PLS-3 than the PLS-4 (personal communications with P. Yoder, October 18, 2004, and C. Wesby, November 2004). It took approximately 50 to 60 min to administer the PLS-4 for the current sample. Second, Suen (2004) argued that the sensitivity reviews conducted by the PLS-4 test developers were not adequate to identify potential ethnically and culturally related bias. Suen suggested that a more formal analysis, differential item functioning analysis (DIF), be performed to determine if test items functioned differently across race/ethnicity. The purpose of the DIF analyses is to determine the quality of each test item and whether the test is measuring the construct in a similar manner in both ethnic groups. To address this concern, Qi and Marley (2009) examined the PLS-4 using DIF to determine whether item bias was present in a large sample of 440 English-speaking Hispanic children and European American preschool children enrolled in Head Start programs. Qi and Marley identified 1 (AC Item 55) of the 61 items on the AC subscale and 2 (EC Items 30 and 31) of the 66 items on the EC subscale displaying DIF. Specifically, AC Item 55 and EC Item 30 appeared to be more difficult for English-speaking Hispanic children than for European American children, whereas the opposite was true for EC Item 31. Second, Suen stated that it is more important to report the degrees of positive predictive power and negative predictive power than the sensitivity and specificity reported in the test manual from a clinical utility perspective.

We used the *Peabody Picture Vocabulary Test–III* (PPVT-III; Dunn & Dunn, 1997) as a comparison test in this study because vocabulary acquisition is a crucial aspect of language development. Words are basic building blocks of meaning. Researchers have suggested that lexical acquisition deficits are a common characteristic of language delays in clinical practice and research (Gray, Plante, Vance, & Henrichsen, 1999). Consequently, deficits in developmentally appropriate lexical acquisition are useful in identifying language delays. The PPVT-III has excellent internal consistency (Cronbach's α > .90). Furthermore, the PPVT-III possesses other desirable characteristics of a good assessment measure. For example, ample data exist supporting the validity of the scores, quality of standardization procedures, reference groups, and cost efficiency (Campbell, Bell, & Keith, 2001; Dunn & Dunn, 1997; Hodapp & Gerken, 1999).

There is a paucity of research related to performance of English-speaking Hispanic preschool children from low-income families on the PLS-4. Because of the widespread adoption of the PLS-4, it is critical that studies be done with children from diverse cultural backgrounds. The present study was designed to examine the psychometric properties of PLS-4 with a sample of English-speaking Hispanic and European American children from low-income families. The specific research questions were as follows:

1. What is the internal consistency of the PLS-4 subscale and total scores?
2. Does the evidence support the concurrent validity of the PLS-4?
3. Using cutoff scores of 70 and 77 (2 and 1.5 *SD*s below the mean) on both measures, what is the agreement between the PLS-4 and the PPVT-III in identification of children with potential language delays?
4. Do the PLS-4 and PPVT-III result in differential prevalence rates of language delays by ethnicity?

Method

Participants

The participants in this study were 440 children (214 girls and 226 boys) who were attending 41 Head Start classrooms in a medium-sized city in the southwestern United States. These participants were selected as part of a larger longitudinal study examining language, behavioral, and social skills of preschool children in Head Start programs. The longitudinal study included three cohorts, with each cohort member having the first wave of data collected in the fall of 2004, 2005, and 2006. The present study used the first-wave data collected on each of the three cohorts. Three hundred seventy seven (85.7%) children were Hispanic and 63 (14.3%) were European Americans, who served as a comparison group. At the beginning of the first wave of data collection, children ranged in age from 34 to 62 months, with a mean age of 48.79 months (*SD* = 6.85). All participants were from families that met the federal guidelines for low-income status. To participate in Head Start, family income must fall below the federal poverty line. Of the mothers who provided information (70%) regarding maternal education level, 28% had 11th grade or less, 33% had high school or GED equivalent, 34% had some college or technical school, and 5% had college degrees.

We selected children on the basis of the following criteria: (a) all children were monolingual English speakers who were not enrolled in bilingual classrooms; (b) children had no obvious or diagnosed hearing, visual, motor, cognitive, or psychiatric deficits; and (c) no Individualized Education Program (IEP) was in place, with the exception of IEPs for speech and language delays. All participants had signed parental consent. Information regarding children's language use was collected from multiple sources. First, the Head Start coordinator for each center identified monolingual English speakers based on enrollment information provided by parents. Second, teachers and research assistants further verified each child's language speaking status. If the child used a Spanish word during testing, the testing was discontinued. Two children were excluded from the study.

Procedures

Children were tested individually with a battery of measures by trained graduate research assistants, all of whom participated in a series of training sessions for approximately 8 hours on the PLS-4 and the PPVT-III by two licensed SLPs and the first author. Prior to collecting data, two fidelity observation sessions were performed with each examiner to ensure proper administration of the measures. Furthermore, the trainers observed the research assistants during the 1st week of testing to ensure that all testing followed standardized procedures.

We administered the PLS-4 individually to each participating child in a quiet area or a private room in each center. It took 50 to 60 min to administer the PLS-4. The two subscales of the PLS-4 were given in counterbalanced order. The PPVT-III (Form A) also was administered individually to each participating child (5-15 min) in the same session or within the same week the PLS-4 was administered. The tester first spent time in the children's classrooms getting familiar with the children and their teachers. Before each test administration, testers talked and played with the individual child. When a child became fatigued or appeared disengaged, testing was discontinued and a water break or play session was provided. After a period of time, when children were rested, testing was resumed until each child completed the test. Each child's responses were scored in accordance with the examination manual. Raw scores were converted to standard scores and percentiles.

Measures

Children's expressive and auditory language skills were assessed with the PLS-4. We chose this test for the present study because of its acceptable psychometric characteristics and its claim to be culturally sensitive to children considered at risk.

The *Peabody Picture Vocabulary Test–III* (PPVT-III; Dunn & Dunn, 1997) is a receptive vocabulary test, which had been shown to be appropriate for use with young children from low-income families in previous research (Qi et al., 2006; Washington & Craig, 1999). The racial and ethnic representation in the standardization sample of the PPVT-III consisted of 64.4% European American, 18.1% African American, 12.9% Hispanic, and 4.6% Other.

Data Analysis

Data analyses proceeded in three stages. First, we conducted preliminary analyses by performing an analysis of variance (ANOVA) to assess gender differences for both Hispanic and European American children for each test. Second, we assessed the internal consistency of the PLS-4 using the Kuder Richardson–20 formula (KR-20; Kuder & Richardson, 1937). Third, we conducted bivariate correlational analyses of the PPVT-III scores in regards to the PLS-4 AC and EC subscales to examine the concurrent validity and predictive invariance of the PLS-4. Next, to further examine the scores for concurrent validity evidence and to ascertain the presence of predictive invariance, we performed a hierarchical regression with the PPVT-III as the dependent variable (Cohen, Cohen, West, & Aiken, 2003) and the PLS-4 subscales as independent variables. We implemented this analysis with the following three steps. First, demographic characteristics were entered into a regression model. Second, the PLS-4 AC and EC subscales standard scores were entered. Finally, the interactions of the demographic variables were entered into the regression equation to test for predictive invariance (Millsap, 1995). In other words, we tested the predictive invariance assumption. The assumption of predictive invariance is best described as a scale having a similar relationship with a criterion, in this case the PPVT-III, across groups (Millsap, 1995). If the relationship between the measure of interest and the criterion differs between groups, the construct validity of the measure of interest is questionable. At each step, we recorded the model R^2 and change in R^2, along with the statistical significance. Finally, we performed a sensitivity and specificity analysis between the PLS-4 and PPVT-III with commonly used cutoff scores. Cases with missing data were deleted listwise for all analyses. The sensitivity/specificity analysis resulted in the following conditional probabilities of interest: sensitivity, the probability of being screened as having a property of interest given that the property is present on the standard; specificity, the probability of being screened as not having the property when the property is not present; positive predictive value (PV^+), the probability of having the property given that the screening instrument indicates the presence of the property; and negative predictive value (PV^-), the probability of not having the property given that the screening instrument indicates the property is present. There is an inverse relationship between sensitivity and specificity. For example, a language diagnostic instrument would have 100% sensitivity if all participants were classified as language delayed. However, the specificity of the test would be 0% in this instance.

Table 1. PLS-4 and PPVT-III Standard Score Means and Standard Deviations by Gender

	Hispanic, *M* (*SD*)			European American, *M* (*SD*)		
Source	Boys	Girls	Total	Boys	Girls	Total
PLS-4 AC	89.47 (12.69)	93.00 (11.76)	91.19 (12.34)	91.28 (15.20)	95.81 (19.84)	93.51 (17.64)
PLS-4 EC	92.49 (13.55)	95.70 (11.71)	94.00 (12.80)	94.38 (14.18)	99.58 (15.59)	96.94 (15.00)
PLS-4 Total	90.23 (13.04)	93.69 (12.22)	91.92 (12.34)	95.20 (11.83)	97.76 (17.98)	94.65 (17.27)
PPVT-III	82.52 (13.56)	84.66 (12.85)	83.56 (13.25)	88.90 (16.55)	88.23 (19.92)	88.57 (18.14)

Note: PLS-4 = *Preschool Language Scale–4*; AC = Auditory Comprehension subscale; EC = Expressive Communication subscale; PPVT-III = *Peabody Picture Vocabulary Test–III*.

Results

Preliminary Analysis

Mean standard score performances for the PLS-4 and PPVT-III are presented in Table 1. In this study, girls scored higher than boys on the PLS-4. Specifically, girls (M = 93.41, SD = 13.21) scored significantly higher than boys (M = 89.73, SD = 13.05) on the PLS-4 AC subscale, F(1, 438) = 8.59, p = .004, Cohen's d = .28. Girls (M = 96.26, SD = 12.38) scored significantly higher than boys (M = 92.76, SD = 13.62) on the EC subscale, F(1, 438) = 7.91, p = .005, Cohen's d = .26; and girls (M = 94.30, SD = 13.27) scored significantly higher than boys (M = 90.88, SD = 12.97) on the PLS-4 Total Language scale, F(1, 438) = 7.83, p = .005, Cohen's d = .26. No ethnic group–related differences were observed on the PLS-4 (all ps > .10).

In contrast, statistically significant differences based on gender were not observed on the PPVT-III, F(1, 425) = 1.64, p = .20; although statistical differences based on ethnic group membership were observed. Because of the violation of the homogeneity of variance assumption, a Welch adjusted degrees of freedom ANOVA was used to compare Hispanic and European American children on the PPVT-III, Welch's F(1, 70.5) = 4.28, p = .04. The results showed that Hispanic children (M = 83.56, SD = 13.25) scored significantly lower on the PPVT-III than did European American children (M = 88.57, SD = 18.14), Cohen's d = .35.

Internal Consistency of the PLS-4

For Hispanic children, the KR-20 values were .90 for the AC subscale, .93 for the EC subscale, and .95 for the total scale. For European American children, the KR-20 values were .92 for the AC subscale, .93 for the EC subscale, and .96 for the total scale. For the entire sample, the KR-20 reliability coefficients of the PLS-4 scales were .90 for the AC subscale, .93 for the EC subscale, and .95 for the total scale. All reliability coefficients were greater than .90 for the entire sample. The PLS-4 score reliabilities were in a range considered "excellent" for the purpose of diagnostic testing (Nunnally & Bernstein, 1994).

Concurrent Validity

To obtain empirical evidence of concurrent validity, we examined correlations between scales that purport to measure similar constructs. The PLS-4 subscales and total scores were moderately correlated with the PPVT-III scores. See Table 2 for stratified zero-order correlations.

The hierarchical regression at Step 1 was statistically significant, F(2, 423) = 4.05, p = .01 and R^2 = .019 (see Table 3). At this step, ethnicity was a significant predictor of children's performance on the PPVT-III (standardized β = –.123, with European American as the reference group). This result indicated the mean performance of Hispanic children was 12% of a standard deviation below that of European American children on the PPVT-III. At Step 2, the PLS-4 AC and EC subscales standard scores were included in the regression. The adding of the two subscale scores accounted for an additional 31% of the variance in the PPVT-III, F(2, 421) = 99.66, p < .001. The total model at this step was also statistically significant, F(4, 421) = 52.81, p < .001 and R^2 = .33, with standardized βs = .31 and .29 for PLS-4 AC and EC subscales, respectively. This finding indicated that a standard deviation change in either PLS-4 subscale resulted in nearly a third of a standard deviation change in the PPVT-III score. The third step introduced the interaction terms between the PLS-4 subscales and the demographic characteristics (gender and ethnicity) to determine whether the PLS-4 subscales homogenously predicted PPVT-III scores across gender and ethnicity. This step was not statistically significant, F(5, 416) = 1.86, p = .10, whereas the total model was statistically significant, F(9, 425) = 24.74, p < .001. The results from the third step provided no evidence of predictive invariance. Therefore, we interpreted the betas in the second step, which suggested that the AC and EC subscale standard scores on the PLS-4 were moderate predictors of performance on the PPVT-III after accounting for variance associated with gender and ethnicity.

Sensitivity and Specificity Analyses

The PLS-4 Examiner's Manual encourages clinicians to establish and follow school-district-based standards for

Table 2. Correlations Between Language Measures for Hispanic, European American Children, and the Entire Sample

	PLS-4 AC	PLS-4 EC	PPVT-III
Hispanic (*n* = 377)			
PLS-4 Total	.93*	.93*	.50*
PLS-4 AC		.75*	.49*
PLS-4 EC			.49*
European American (*n* = 63)			
PLS-4 Total	.95*	.94*	.71*
PLS-4 AC		.85*	.66*
PLS-4 EC			.66*
Entire sample (*N* = 440)			
PLS-4 Total	.93*	.93*	.55*
PLS-4 AC		.77*	.54*
PLS-4 EC			.53*

Note: PLS-4 = *Preschool Language Scale–4*; AC = Auditory Comprehension subscale; EC = Expressive Communication subscale; PPVT-III = *Peabody Picture Vocabulary Test–III*.
*$p < .05$.

Table 3. Hierarchical Regression of the PPVT-III

Explanatory variable	Model 1	Model 2	Model 3
Demographics			
Gender (ref = males)	.060	−.014	.052
Ethnicity (ref = European American)	−.123*	−.088*	.366
PLS-4 subscales			
AC subscale		.313**	.230
EC subscale		.289**	.549*
Interactions			
Ethnicity by gender			.176
Gender by AC scores			.706
Gender by EC scores			−.927
Race by AC scores			−.147
Race by EC scores			−.393
R^2 change		.315**	.015
R^2	.019*	.334**	.349**

Note: Standardized beta coefficients reported. PLS-4 = *Preschool Language Scale–4*; AC = Auditory Comprehension subscale; EC = Expressive Communication subscale; PPVT-III = *Peabody Picture Vocabulary Test–III*.
*$p < .05$. **$p < .001$.

"language disorder" (Zimmerman et al., 2002, p. 124). For this study, we examined cutoff scores of 70 and 77 on both measures for cross comparison between the PLS-4 and the PPVT-III, because 2 *SD*s and 1.5 *SD*s below the mean are commonly used benchmarks in language delay diagnosis (Leonard, 1998; Qi et al., 2003; Qi et al., 2006).

In the first sensitivity/specificity analysis, we used cutoff scores of 70 on both the PLS-4 and PPVT-III (see Table 4). In this circumstance, the sensitivity and specificity of the PLS-4 in relationship to the PPVT-III were 26.0% and 97.9%, respectively. We further calculated predictive values of the PLS-4 to determine whether they could detect a potential language delay. The PV$^+$ of the PLS-4 was 68%, which means that 68% (17 of 25) of the children identified as having a potential language delay by the PLS-4 would be considered as having a potential language delay by the PPVT-III. The PV$^-$ of the PLS-4 was 88% in this sample; this indicates that 88% (367 of 415) of the children identified as having typical language by the PLS-4 would be similarly assessed by the PPVT-III. It should be noted that in this sample 17% of the children identified by the PLS-4 as having typical language would be considered as having a potential language delay by the PPVT-III. In other words, 17% of children identified as having typical language by the PLS-4 would be false negatives. The overall accuracy of the PLS-4 for identification of potential language delays was 87% (i.e., the percentage of true positives plus true negatives). Furthermore, 5.0% (19) of the Hispanic and 9.5% (6) of the European American children would be considered for further clinical evaluation for language delays.

In the second analysis, we examined the cutoff score of 77. A cutoff score of 77 on both measures resulted in a sensitivity of 29.0% and a specificity of 93.4% (see Table 5). As previously described, increases in sensitivity results in a decrement in specificity. The PV$^+$ of the PLS-4 was 58%, which means 58% (31 of 53) children identified as having a potential language delay by the PLS-4 were also considered as having a potential language delay by the PPVT-III. The PV$^-$ of the PLS-4 was 80% in this sample; this indicated that 80% (311 of 387) of the children identified as having typical language by the PLS-4 would be similarly assessed by the PPVT-III. This would result in 20% of children identified as having typical language by the PLS-4 being considered false negatives by the PPVT-III. The overall accuracy of the PLS-4 for identification of potential language delays using the cutoff score of 77 was 78%. In addition, 11.7% (44) and 14% (9) of Hispanic and European American children would be identified for further evaluation, respectively.

The sensitivity of the PLS-4 in relation to the PPVT-III is quite low and both cutoff scores would result in an unacceptable number of false negatives. However, comparing the two cutoff scores, the cutoff score of 77 might be slightly better than that of 70 in PLS-4 in correctly identifying potential language delays (i.e., 29% vs. 26%).

Prevalence Rates

The commonly used cutoff scores on PPVT-III resulted in a greater likelihood of identifying children as suspect for having a potential language delay relative to the same cutoff scores on the PLS-4. In our sample, 15% and 24% of

Table 4. Sensitivity and Specificity Analysis of Language Delay Using Cutoff Scores of 70

	Delay status based on PPVT-III		
Identified by PLS-4 as having a language delay	Children with a language delay	Typically developing language	Row total
Positive	17	8	25
Negative	48	367	415
Column total	65	375	440

Note: Sensitivity = 17/65 = .26; positive predictive value (PV$^+$) = 17/25 = 68%; specificity = 367/375 = .97; negative predictive value (PV$^-$) = 367/415 = 88%; overall accuracy = (17 + 367)/440 = 87%. PPVT-III = *Peabody Picture Vocabulary Test–III*; PLS-4 = *Preschool Language Scale–4*.

Table 5. Sensitivity and Specificity Analysis of Language Using Cutoff Scores of 77

	Delay status based on PPVT-III		
Identified by PLS-4 as having a language delay	Children with a language delay	Typically developing language	Row total
Positive	31	22	53
Negative	76	311	387
Column total	107	333	440

Note: Sensitivity = 31/107 = .29; positive predictive value (PV$^+$) = 31/53 = 58%; specificity = 311/333 = .93; negative predictive value (PV$^-$) = 311/387 = 80%; overall accuracy = (31 + 311)/440 = 78%. PPVT-III = *Peabody Picture Vocabulary Test–III*; PLS-4 = *Preschool Language Scale–4*.

Hispanic children would be identified as requiring further speech and language evaluation using cutoff scores of 70 and 77 on the PPVT-III, respectively. For European American children, a similar pattern emerged on the PPVT-III, with 13% identified at cutoff scores of 70 and 20% identified at cutoff scores of 77. On the other hand, the PLS-4 resulted not only in a lower rate of identification at the same cutoff scores; it also identified European Americans at a slightly greater rate relative to Hispanic children. When the PLS-4 cutoff score was set to 70, 5.0% of Hispanic children and 9.5% of European American children were identified as needing further evaluation. At a cutoff score of 77, the pattern held with 11% of Hispanic and 14% of European American children being identified for further evaluation.

Discussion

In this study, our main goal was to evaluate the psychometric properties of the PLS-4 for English-speaking Hispanic and European American preschool children who attended Head Start. There were five major findings.

First, the observed gender differences replicated results of previous studies that have found that girls from low-income families tended to perform higher than did boys from low-income families on standardized language tests (Qi et al., 2003). Furthermore, we found no statistical differences between Hispanic and European American children on the PLS-4 scales. However, we observed statistical differences on the PPVT-III, with European American children achieving higher scores on average relative to Hispanic children (Cohen's d = .35). This finding, when considered along with the related greater rate of language delay identification of Hispanic children on the PPVT-III relative to the PLS-4, indicates that the PLS-4 might be considered a promising instrument for use with English-speaking Hispanic children from low-income families. In addition, the lower mean performance of Hispanic children on the PPVT-III might result in a greater number of Hispanic children being identified for further comprehensive evaluation and possible special education and speech and language services. This key finding should be investigated further to determine whether other ethnic differences are present in identification rates between the PLS-4 and PPVT-III.

Second, the PLS-4 subscales and total scale scores had excellent internal consistency. All KR-20 values for AC, EC, and total scale of the PLS-4 were between .90 and .95, regardless of ethnic background. These reliabilities would be considered "excellent" for usage in diagnostic situations (Kline, 1993; Nunnally & Bernstein, 1994). This finding suggests that the scores generated by the PLS-4 are reliable and that the standard errors of measurement will be low for individual scores. For example, with the lowest observed reliability of .90 and the standard errors of measurement of 1.96 (i.e., a 95% confidence interval), a child who scores 70 on the PLS-4 total scale would be expected to have a true score between 69.39 and 70.61.

Third, the concurrent validity of the PLS-4 indicates that PLS-4 subscale and total scores are moderately related to the PPVT-III scores. This finding provides criterion-related validity evidence for the PLS-4. The regression of the PPVT-III on the PLS-4 subscales echoes the results of the zero-order correlations of the PLS-4 and PPVT-III in support of criterion-related validity. The lack of statistical interactions between the demographic characteristics and the PLS-4 AC and EC subscales is supportive of the assumption of predictive invariance. However, the statistical power to identify interactions is limited in the current sample because of the small number of European American children. Further research with larger samples is necessary to identify meaningful statistical interactions if they are present.

Fourth, sensitivity and specificity analysis comparing commonly used cutoff scores of 70 and 77 revealed the accuracy of the PLS-4 identifying children with language delays as having language delays (sensitivity) was 26% and 29%, respectively. In both cut score scenarios the sensitivity of the PLS-4 was low in relationship to the PPVT-III using the criteria recommended by Plante and Vance (1994) of 80%-89% sensitivity for an evaluation of "fair." However, using the cutoff scores of 70 and 77, the accuracy of the PLS-4 identifying children with typical language development as having typically developing language (specificity) was 97% and 93%, which were higher than recommended criterion of 90%-100% of "good" (Plante & Vance, 1994). The overall accuracy between the PLS-4 and PPVT-III was greatest when a cutoff score of 70 was used on both measures. A cutoff score of 70 resulted in an overall accuracy of 87% while a cutoff score of 77 resulted in an overall accuracy of 78%. However, it should be noted the improved accuracy associated with setting the cutoff score at 70 was due to increased specificity, not sensitivity. The relatively lower sensitivity associated with cut scores of 70 could potentially result in children with true language delays failing to receive further evaluation.

Finally, with English-speaking Hispanic children, the PPVT-III identified potential language delays at a higher rate than did the PLS-4. Using cutoff scores of 70 and 77 on the PPVT-III, respectively, 15% and 24% of Hispanic and 13% and 20% of European American children would be identified as having a potential language delay. The PLS-4 was more conservative in the identification of children with potential language delays, with 5% and 11% for Hispanics and 9.5% and 14% for European American children using cutoff scores of 70 and 77, respectively.

We expected that the PPVT-III would estimate a lower prevalence of language delays because it addresses vocabulary only, whereas the PLS-4 assesses semantics (content) through tasks that focus on both vocabulary and concept, structure (form) through tasks focusing on syntax and morphology, and other skills such as integrative language skills and phonological awareness (Zimmerman et al., 2002). Children with language disorders are expected to have impairments in these areas because language semantics and structure are interrelated components that represent children's receptive and expressive language (Lahey, 1988). On the basis of this rationale, we expected that the PLS-4 would result in a greater number of children being identified as requiring further evaluation for potential language delay. However, our findings indicated that the PPVT-III identified a greater number of children who might be at risk for potential language delay in the present sample. This is consistent with Andersson and Pitti's (2003) results that the *Fluharty Preschool Speech and Language Screening Test–Second Edition* (FLUHARTY-2; Fluharty, 2001) identified more children in Head Start programs as requiring further language testing than did the PLS-4. They found that even combined with a teacher questionnaire, the PLS-4 (using either of the recommended criteria) identified only children with more severe/global difficulties.

Strengths, Limitations, and Recommendations for Future Research

There are two notable strengths of the study. The requirement that children's family income be below a certain level for eligibility in a Head Start program provides a control for income. To a degree, it can be assumed that these children come from a comparable economic background. This feature reduces the likelihood that observed differences between Hispanic and European American children are due to differences in socioeconomic status (SES). This finding provides further preliminary evidence of the PLS-4 being an appropriate measure for identifying potential language delays in preschool children from low-income families (Qi & Marley, 2009). The second strength, in an applied sense, is that the sample was gathered from urban Head Start centers serving low-income families. Our results provide preliminary findings on the reliability and validity of the PLS-4 with a large sample of English-speaking Hispanic children from low-income families.

There are two limitations of the study. First, since all measures are fallible, there is not a gold standard in the clinical sense for the identification of language delay. Therefore, the relationship of one fallible measure to another fallible measure is not strong evidence of accuracy. All that can be done is to relate several measures to one another, resulting in the uncomfortable conclusion that the measures correlate with one another, indicating something systematic is being assessed. Another potential limitation is that the sample size of the comparison group was very small. In addition, the nature of the sample may limit the generalizability of the findings to comparable populations. In summary, findings from this exploratory study must be interpreted with caution, as mentioned earlier. Our findings would be further supported by evaluating the PLS-4 in relation to other indicators, such as clinical judgment. Future research focusing on examining the reliability and validity of the PLS-4 would benefit from the inclusion of children in geographically different regions, as well as children from middle- or higher-SES families and different racial and ethnic backgrounds.

Clinical Implications

The results of our study highlight an area that deserves particular attention in additional research on language assessment with preschool populations from low-income families. Although our study provides some preliminary evidence that the PLS-4 appears to be a promising language instrument for assessing preschool children from low-income families, it is worth noting that SLPs and other related professionals should use the PLS-4 with caution because of the possibility that it may underidentify children with potential language delays. Two special considerations most SLPs have to address before selecting an instrument for screening and identification purposes are (a) weighing measure sensitivity and specificity and (b) deciding on cutoff scores (McCauley, 2001). In this study, we found that PLS-4 has relatively low sensitivity and high false-negative rates with cutoff scores of 70 and 77. With false-negative rates of 12% and 20%, the chance of a child with a potential language disorder not being identified is likely to be unacceptably high. Our finding also suggests that the overall accuracy of the PLS-4 in identifying a potential language delay is higher when a cutoff score of 70 is used compared with a cutoff score of 77. Plante and Vance (1995) suggested that SLPs should also take the specific testing situation into consideration when evaluating a measure's sensitivity and specificity. For example, in early childhood education settings like Head Start programs, a measure's low sensitivity might be considered acceptable as a teacher can still refer to a child if he or she has concerns about the child's language skills even if the child may have passed the screening test before. Clinicians should also consider the predictive values in the context of the prevalence rates for language delay (Maxwell & Satake, 2006). Normally, when language delay occurs less frequently, such as in a community sample instead of a clinic sample, the test might result in a low PV^+ and a high PV^-. In summary, assessment best practice should combine the use of standardized tests with other informal language assessment as well as clinical judgment. The combination of assessments can provide a general picture of children's language abilities when compared with their same-aged peers.

Acknowledgment

We thank the children and families that participated in this study for contributing their time as well as the teachers and directors of the Head Start centers for their collaboration. In addition, we wish to thank the many research assistants for their help with data collection and data entry. Thanks are also extended to Paul Yoder at Vanderbilt University for his helpful advice and insightful input into this manuscript.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research and/or authorship of this article: This research was supported by National Institute on Deafness and Other Communication Disorders (NIDCD) Grant 1RO3 (DC006117-01A2 and a Large Research Allocation Committee (RAC) grant from the University of New Mexico to the first author.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Andersson, L., & Pitti, M. (2003, November). *Assessing language in Head Start children using teacher report and test scores*. Poster session presented at the 2003 annual convention of the American Speech Language Hearing Association, Chicago, IL.

Campbell, J. M., Bell, S. K., & Keith, L. K. (2001). Concurrent validity of the Peabody Picture Vocabulary Test–Third Edition as an intelligence and achievement screener for low SES African American children. *Assessment, 8,* 85–94.

Champion, T. B., Hyter, Y. D., McCabe, A., & Bland-Stewart, L. M. (2003). A matter of vocabulary: Performances of low-income African American Head Start children on the Peabody Picture Vocabulary Test-III. *Communication Disorders Quarterly, 24*, 121–128.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test—Third Edition*. Circle Pines, MN: American Guidance Service.

Fluharty, N. B. (2001). *Fluharty Preschool Speech and Language Screening Test* (2nd ed.). Austin, TX: Pro-Ed.

Frankenburg, W. K., Dodds, J., Archer, P., Bresnick, B., Maschka, P., Edelman, N., & Shapiro, H. (1990). *DENVER II*. Denver, CO: Denver Developmental Materials.

Garciá, E., & Jensen, B. (2009). Early educational opportunities for children of Hispanic origins. *Social Policy Report: Giving Child and Youth Development Knowledge Away, 23*(2), 3–19.

Gray, S., Plante, E., Vance, R., & Henrichsen, M. (1999). The diagnostic accuracy of four vocabulary tests administered to preschool-age children. *Language, Speech, and Hearing Services in Schools, 30*, 196–206.

Hegde, M. N., & Maul, C. A. (2006). *Language disorders in children: An evidence-based approach to assessment and treatment*. Boston, MA: Allyn & Bacon.

Hodapp, A. F., & Gerken, K. C. (1999). Correlations between scores for Peabody Picture Vocabulary Test-III and the Wechsler Intelligence Scale for Children-III. *Psychological Reports, 84*, 1139–1142.

Huang, R., Hopkins, J., & Nippold, M. A. (1997). Satisfaction with standardized language testing. *Language, Speech, and Hearing Services in Schools, 28*, 12–23.

Kaderavek, J. N. (2011). *Language disorders in children: Fundamental concepts of assessment and intervention*. Boston, MA: Allyn & Bacon.

Kline, P. (1993). *The handbook of psychological testing*. London, UK: Routledge.

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*, 151–160.

Lahey, M. (1988). *Language disorders and language development*. New York, NY: Macmillan.

Leonard, L. (1998). *Children with specific language impairment*. Cambridge, MA: MIT Press.

Maxwell, D. L., & Satake, E. (2006). *Research and statistical methods in communication sciences and disorders*. Clifton Park, NY: Thomson Delmar Learning.

McCauley, R. J. (2001). *Assessment of language disorders in children*. Mahwah, NJ: Lawrence Erlbaum.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*, 577–605.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3rd ed.). New York, NY: McGraw-Hill.

Plante, E., & Vance, R. (1994). Selection of preschool speech and language tests: A data-based approach. *Language, Speech, and Hearing Services in Schools, 25*, 15–23.

Plante, E., & Vance, R. (1995). Diagnostic accuracy of two tests of preschool language. *American Journal of Speech-Language Pathology, 4*, 70–76.

Qi, C. H., Kaiser, A. P., Milan, S., & Hancock, T. (2006). Language performance of low-income African American children on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools, 37*, 1–12.

Qi, C. H., Kaiser, A. P., Milan, S., Yzquierdo, Z., & Hancock, T. (2003). The performance of low-income African American children on the Preschool Language Scales-3. *Journal of Speech, Language, and Hearing Research, 43*, 576–590.

Qi, C. H., & Marley, S. C. (2009). Differential item functioning analysis of the Preschool Language Scale-4 between English-speaking Hispanic and European American children from low-income families. *Topics in Early Childhood Special Education, 29*, 171–180.

Restrepo, M. A., Schwanenflugel, P., Blake, J., Neuhart-Pritchett, S., Cramer, S., & Ruston, H. P. (2006). Performance on the PPVT-III and the EVT: Applicability of the measures with African American and European American preschool children. *Language, Speech, and Hearing Services in Schools, 37*, 17–27.

Stanton-Chapman, T. L., Chapman, D. A., Kaiser, A. P., & Hancock, T. B. (2004). Cumulative risk and low-income children's language development. *Topics in Early Childhood Special Education, 24*, 227–237.

Suen, H. K. (2004). Suen, H. K. (2005). *Review of the Preschool Language Scale, 4th edition*. In R. A. Spies & B. Blake (Eds.), *The sixteenth mental measurements yearbook* (pp. 831–834). Lincoln, NE: Buros Institute of Mental Measurement.

The Head Start Family and Child Experiences Survey. (2008). Beginning Head Start: Children, Families and Programs in Fall 2006. Retrieved from http://www.acf.hhs.gov/programs/opre/hs/faces/reports/beginning_hs/beginning_hs.pdf

Washington, J., & Craig, H. (1999). Performances of at-risk African American preschoolers on the Peabody Picture Vocabulary Test-III. *Language, Speech, and Hearing Services in Schools, 30*, 75–82.

Zimmerman, I. L., & Castilleja, N. F. (2005). The role of a language scale for infant and preschool assessment. *Mental Retardation and Developmental Disabilities Research Reviews, 11*, 238–246.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (1992). *Preschool Language Scale, Third Edition*. San Antonio, TX: Psychological Corporation.

Zimmerman, I. L., Steiner, V. G., & Pond, R. E. (2002). *Preschool Language Scale, Fourth Edition*. San Antonio, TX: Psychological Corporation.

About the Authors

Cathy H. Qi, PhD, is an associate professor of special education at the University of New Mexico. Her current interests include language and behavior assessment and autism spectrum disorders.

Scott C. Marley, PhD, is an assistant professor of educational psychology at the University of New Mexico. His interests include research methodology, applied statistics, Native American education, and reading strategies.