LSHSS

Performance on the PPVT–III and the EVT: Applicability of the Measures With African American and European American Preschool Children

María Adelaida Restrepo
Arizona State University, Tempe
Paula J. Schwanenflugel
Jamilia Blake
Stacey Neuharth-Pritchett
Stephen E. Cramer
Hilary P. Ruston
The University of Georgia, Athens

Research indicates that language acquisition does not differ for African American (AA) and European American (EUA) children (Huston, McLoyd, & Garcia Coll, 1994; Mount-Weitz, 1996; Roberts, Bruchinal, & Durham, 1999), yet performance differences between these two groups on standardized language assessments have often been noted (Champion, Hyter, McCabe, & Bland-Stewart, 2003; Fagundes, Haynes, Haak, & Moran, 1998; Hammer, Penncock-Roman, Rzasa, & Tomblin, 2002; Qi, Kaiser, Milan, Yzquierdo, & Hancock, 2003; Washington & Craig, 1992). On average, AA children tend to score 1 *SD* below EUA peers on standardized language and cognitive measures (Brooks-Gunn, Klebanov, & Duncan, 1996; Hammer et al., 2002; Kaufman & Kaufman, 1983; Mercer, 1979; Reynolds, Lowe, & Saenz, 1999; Whitehurst, 1997). Of particular interest is AA children's performance on vocabulary measures (Champion et al., 2003; Hammer et al., 2002; Qi et al., 2003). Although Reynolds et al. (1999) argued that clinicians should anticipate mean differences between ethnic groups, most researchers would agree that performance differences of ≥1 *SD*s warrant caution in the interpretation

ABSTRACT: **Purpose:** The purpose of this study was to determine whether two vocabulary measures were appropriate for the evaluation of African American children and children whose mothers have low education levels, regardless of gender.
Method: Data were collected for 210 high-risk, preschool children from a southeastern state in the United States on the Peabody Picture Vocabulary Test—Third Edition (PPVT–III; L. M. Dunn & L.M. Dunn, 1997) and the Expressive Vocabulary Test (EVT; K. T. Williams, 1997).
Results: Results indicated that African American children and children whose mothers had low education levels tended to score lower on both measures than did children from European American backgrounds and children whose mothers had a high school or higher education; however, this effect was larger for the PPVT–III.
Clinical Implications: Data suggest that the EVT is a better indicator of a child's "vocabulary" skill, and that the PPVT–III has a greater tendency than the EVT to place African American children and children whose mothers have low education levels at risk for being unfairly identified as presenting with a potential language disorder. These data indicate that practitioners should use alternative assessment methods such as nonstandard and dynamic assessments to test children's vocabulary skill. In particular, if they use the PPVT–III, practitioners should take great caution in interpreting test results as evidence of a vocabulary problem in African American children and children whose mothers have low education levels.

KEY WORDS: vocabulary, African American, maternal education, differential item functioning, test bias, PPVT, EVT

of the instrument for the discrepant group (Champion et al., 2003; Qi et al., 2003).

Possible Sources of Differences in Performance Between AA and EUA Children

Performance differences between AA and EUA children in language assessment and vocabulary assessment may be due partially to the mismatch between AA culture such as language socialization practices and the methods in which standardized tests require children to demonstrate their knowledge. For example, Fagundes et al. (1998) found that AA children performed better on the Preschool Language Instrument (Blank, Rose, & Berlin, 1978) when a physical context and a thematic activity were provided than when the test was given with no contextual support. Others have suggested that there is a mismatch between the number and variety of question types to which AA and EUA children have been acculturated compared to those represented by many language and vocabulary tests (Heath, 1983). For example, Washington (2001) reported that the number and types of questions that AA parents ask vary during play interactions with preschool children.This would account for why AA children experience more difficulty than their EUA peers with tasks requiring familiarity with questions (e.g., standardized tests). Similarly, Heath found that AA mothers in the rural South infrequently asked single-word response questions but asked more analogy comparisons, explanations, and nonverbal response type of questions. Moreover, Anderson-Yokel and Haynes (1994) found that AA mothers asked significantly fewer yes/no and *wh*- questions than EUA mothers of preschool children during a shared storybook reading activity. In this study, AA children produced more spontaneous verbalizations than responses to questions than did EUA children (Anderson-Yokel & Haynes, 1994).

Children Raised in Poverty

Poverty may also contribute to the lower performance that some young children display on standardized language assessments (Campbell et al., 2001; Kamphaus, 2001; Washington & Craig, 1999). Investigators have identified socioeconomic status (SES) and maternal education level (a proxy for SES) as significant predictors of children's performance on standardized measures and overall intellectual functioning (e.g., Brooks-Gunn et al., 1996; Washington & Craig, 1999). Whitehurst (1997), in a review of his research on the language of children who have been raised in poverty, concluded that these children tend to score ≥1 *SD* below the mean on receptive vocabulary measures such as the Peabody Picture Vocabulary Test—Third Edition (PPVT–III; Dunn & Dunn, 1997), expressive vocabulary, metalinguistic skills, narrative skills, and sentence complexity than their peers with higher SES. In addition, other investigators have documented that the amount of talk that children who have been raised in poverty hear is much less compared to that heard by children who have been raised in middle-class homes (Hart & Risely, 1995; Whitehurst, 1997), thus contributing to low scores on language measures such as vocabulary tests. Hart and Risely, for example, found that word and sentence variability, and the sheer amount of child-directed speech, varied by SES, as determined by parents' occupations. Parents of 1- and 2-year-old children from low-income environments used fewer words and syntactic variations than did their professional and working class counterparts. Hart and Risely argued that the quality and quantity of language that was used in the homes produced the social class variations that were found among children in expressive vocabulary and language development.

Purcell-Gates (1996) indicated that low-SES parents tended to do more shared reading and explicit literacy instruction only after children entered elementary school and received direct literacy instruction. Therefore, limited experiences with literacy in preschool may lead children to have difficulty in responding to decontextualized tasks such as those seen in vocabulary measures that require them to listen to an adult and to point to pictures. Purcell-Gates also found that children from low-SES backgrounds participated in literacy experiences as part of the contextual focus of oral discourse rather than as a separate activity in which literacy, listening, and pointing to pictures is the whole focus of the activity. These studies, therefore, suggest that children with less direct literacy experiences in preschool years and less talk may be at risk of scoring low on decontextualized vocabulary measures, even when they do not present with language disorders.

AA children's performance on standardized tests of linguistic ability is complicated by the interaction between ethnicity, SES, education, and culture. Researchers are unsure whether standardized tests exhibit bias toward AA children because they fail to take into account the interplay of SES and language socialization practices on test performance, or if standardized tests are measuring "true" linguistic ability differences between groups. However, many test developers are sensitive to performance differences and have attempted to ensure the utility and validity of their instruments with all populations. Thus, AA children's lower scores on standardized measures may be attributed to the larger number of AA children living in poverty (Washington, 2001). Nevertheless, it is possible that performance differences in favor of EUA children may still emerge when AA and EUA children from middle-class homes are compared (Washington, 2001).

Peabody Picture Vocabulary Test

The PPVT–III (Dunn & Dunn, 1997) is of particular interest because of the general belief regarding its relative validity for assessment of verbal ability in AA children. For example, the PPVT–III is used in large-scale federally funded preschool research projects, such as the Family and Child Experience Study (FACES; U.S. Department of Health and Human Services [U.S. DHHS], 2003), Even Start programs (U.S. Department of Education, 2004b), Early Reading First (U.S. Department of Education, 2004a), and The Early Childhood Longitudinal Study (National Center for Education Statistics, 1999). The implication of using the PPVT–III in federally funded grants is that, on average, researchers believe that it is a relatively valid measure of children's verbal ability. However, if it is not valid for some groups of children, evaluation of programs might be compromised by use of the PPVT–III, and the policy implications drawn might be problematic for AA children.

Many speech-language pathologists (SLPs) use the PPVT–III as a screening instrument for verbal ability and to evaluate receptive vocabulary, although researchers have warned that this is not an appropriate use of the measure (e.g., Gray, Plante, Vance, & Henrichsen, 1999). The PPVT–III, is preferred by practitioners due to its brevity, ease of administration and scoring,

correlation with intelligence scales (Campbell, 1998; Williams & Wang, 1997), and some limited evidence that it may not be biased against AA children (Washington & Craig, 1999). However, the PPVT–III and earlier versions of the test have sparked controversy surrounding their appropriateness for non-EUA populations (Campbell, Bell, & Keith, 2001; Champion et al., 2003; Goff & Montague, 1980; Washington & Craig, 1992, 1999).

Studies examining the validity of the PPVT–III for AA children have produced conflicting findings. For example, Williams and Wang (1997) examined item bias toward AA children in the PPVT–III standardization sample. Results of their study indicated good item discrimination, with no items warranting deletion. Thus, the item analyses suggested that the PPVT–III did not demonstrate item bias and was appropriate for assessing AA children. Further, Washington and Craig (1999) examined the performance of 59 (25 boys and 34 girls) AA children attending a state-sponsored preschool program for at-risk children (based on income, family density, family histories, and single-parent households). The authors found that AA children's scores did not differ significantly from those of the standardization sample ($M = 91$, $SD = 11$), except for children whose mothers had less than a high school education ($M = 77.3$, $SD = 10.7$). Although maternal education level was noted as a significant predictor of performance for AA children, Washington and Craig concluded that the PPVT–III was adequate for urban AA preschoolers.

Despite results indicating that the PPVT–III may be unbiased, Stockman (2000) questioned the validity of such findings. She examined the differences between the Peabody Picture Vocabulary Test—Revised (PPVT–R; Dunn & Dunn, 1981) and the PPVT–III and found that there were significant changes in item difficulty between revisions. Specifically, the PPVT–III produced higher raw score conversion scores than the PPVT–R. Stockman attributed these differences to changes of word types and frequencies, demographics of the standardization sample, and alterations of the picture stimuli between revisions. She concluded that differences between revisions might result in artificial indexes of improvement or inaccurate diagnosis and educational placement when standard scores between the PPVT–R and the PPVT–III are compared. Further, she urged caution in the use of the PPVT–III with preschool-age children and suggested that the PPVT–III may not be an unbiased test for all children. That is, she stated that more research is needed with other ethnic groups in studies of test bias, that ethnic group performances need to be evaluated using measures of semantic knowledge rather than IQ measures, and that ethnic group analyses based on the identification rate of language delay are still needed to positively conclude that the PPVT–III is not biased for all children.

Ukrainetz and Duncan (2000) argued that the PPVT–III scores may overestimate children's receptive vocabulary. Further, Ukrainetz and Bloomquist (2002) found that the PPVT–III mean for a sample of mostly EUA children with a wide range of SES backgrounds was higher (107) compared to the test mean (100). Concerns of overestimating scores when using the PPVT–III contrast with Stockman's concern that the test may be biased against AA preschool children (2000). Campbell et al. (2001) examined the performance of 416 AA kindergarten children (M age = 6;3 [years;months]) from single-parent homes and low-income homes. They found that the AA children (2000) scored 1 *SD* below the mean of the PPVT–III. As a result, the authors urged caution in the use of the PPVT–III for low-SES AA children.

Champion et al. (2003), similar to Campbell et al. (2001), found that AA children in Head Start programs performed at 1 *SD* below the mean on the PPVT–III; however, they warned that Head Start children, in general, are scoring this low. Nevertheless, they checked the 75 items that children tended to miss the most on the PPVT–III within the AA community and found alternate meanings for 11 of them, suggesting possible item bias, at least for the preschool children. These results are consistent with those of Whitehurst and colleagues, who found that children who were raised in poverty and children in Head Start programs scored below 1 *SD* from the mean, regardless of their ethnic background (see Whitehurst, 1997 for a review).

The findings of the studies described above leave researchers and practitioners with equivocal results concerning the validity of the PPVT–III for evaluating young AA children. Williams and Wang (1997) and Washington and Craig (1999) found the PPVT–III to be unbiased; Ukrainetz and Duncan (2000) argued that it overestimates scores in general; and Campbell et al. (2001), Champion et al. (2003), and Stockman (2000) proposed that the measure was biased against some AA children. Confounding variables of race and ethnic background, maternal education, and SES further complicate the interpretation of these studies. One commonality across studies is their failure to include groups of AA children with a range of maternal education levels and family income. Including comparison groups would allow researchers to test whether the PPVT–III is biased against AA children in general or those with low maternal education levels. Given the ubiquity of the use of the test in large federal evaluation studies and the conflicting findings regarding issues of bias, it is important to determine further whether the test is, indeed, biased. This concern applies to the co-normed, but less researched, measure, the Expressive Vocabulary Test (Williams, 1997), because they can be used together to assess vocabulary.

Expressive Vocabulary Test

The EVT (Williams, 1997) is a measure of children's expressive vocabulary that complements the PPVT–III to provide a broad assessment of children's one-word vocabulary. There is little independent empirical evidence demonstrating the psychometric adequacy of the EVT in regard to its validity for non-EUA populations. Lack of evidence supporting the predictive validity of the EVT is of particular concern because expressive difficulties tend to be more common than receptive ones in young children (Gray et al., 1999), and the EVT and PPVT–III were co-normed, which allows comparisons between tests because the normative sample was the same. Given the concerns of bias with the PPVT, it is therefore imperative to determine whether the same issues apply to its co-normed test, especially when group comparisons are made with the two tests. The use of the EVT, however, does not have the history that the PPVT has, although its use is increasing because of the comparability of the two tests. Co-norming of the tests indicates that the same sample was used to determine the norms of each measure, and thus it is possible to compare a child's performance on both measures.

Studies directly comparing the PPVT–III and EVT are scarce. Gray et al. (1999) investigated whether several vocabulary measures, including the PPVT–III and EVT, were valid for diagnosing children with language impairments. They found that most children who were developing typically obtained

significantly higher PPVT–III scores than EVT scores; however, more children with language impairments obtained significantly higher EVT scores than PPVT–III scores. Gray et al. also noted that there were significant effects for gender on the PPVT–III but not on the EVT, although the direction of the difference was not reported. No effects for minority children (AA or Hispanic) on either measure were found, although the sample was possibly too small to have sufficient power to detect differences. Investigations in which the PPVT–III and EVT were administered as part of a battery of tests suggest that both instruments appropriately measure verbal ability (Gray et al., 1999; Qi et al., 2003; Ukrainetz & Bloomquist, 2002). However, in these studies, the EVT and PPVT–III were used as criterion-related validity measures to validate different language scales. Thus, further studies examining their appropriateness with preschool children of varying ethnicities and maternal education levels are needed, especially when such measures are used for the diagnosis of language disorders.

Sources of Test Bias

To determine whether the PPVT and EVT are biased against certain groups, sources of test bias need to be identified. Brown, Reynolds, and Whitaker (1999) defined test bias as test score differences between cultural or ethnic groups that do not reflect real differences in ability but rather are the result of "problems in the construction, design, administration, or interpretation of tests" (p. 209). Skiba, Knesting, and Bush (2002) discussed a number of possible sources of test bias: problems in construct validity, underrepresentation in the sample, bias in predictive validity, item bias, and language and examiner bias.

Construct validity problems would be identified if groups differ on the factor components of a test or if specific items present problems for a specific group (Reynolds, 2000; Skiba et al., 2000). Bias from sampling may come from the underrepresentation of minorities in a sample, which in turn would bias item selection, and therefore the ethnic group may not have a significant impact on the test. Hickman and Reynolds (1987) however, have questioned this source of bias because they have found no evidence that this is the case. A test would also be biased if it consistently over- or underpredicted a group's performance.

Item bias reflects whether a group has less exposure to process information required by the test, and thus, the group would have difficulties with specific items. The best ways to evaluate item bias are through procedures such as item response theory or differential item functioning (DIF) (Reynolds, 2000). Bias is, however, a poorly understood and emotionally loaded term. Osterlind (1983) suggested that "bias is not the mere presence of a score difference between two groups" (p. 12). Differences in scores or in proportions getting an item correct are more accurately referred to as adverse impact. True bias, which at the item level is usually termed DIF, is present when the probability of getting an item correct differs across groups *who have the same level of latent ability*. Bias may also come from the language or dialect of the examinee and the examiner. Taylor and Lee (1987) argued that because psychological and language tests are often done in a standard dialect, tests on children who use nonstandard dialects may measure the extent of knowledge with the standard dialect rather than true aptitude.

Ultimately, however, appropriate test use falls on the clinicians who must be informed about tests, test norms, and the issues that impact a child's performance on tests, such as educational history, maternal education level, or cultural background. Consequently, clinicians should be aware if a test is biased or has adverse effects against an educational, economic, gender, or ethnic group and any interactions of these. If a test is not appropriate for identifying language disorders in a specific group, it is the use of the test with that group that is problematic and not necessarily the test.

The purpose of the present study was to investigate the validity of the PPVT–III and the EVT for assessing vocabulary in preschool children. The study evaluated whether the PPVT–III and EVT are unbiased for children regardless of ethnicity, gender, and maternal education levels. Item bias and mean group differences across the different categories were used to partially examine whether these two measures are biased. It is hypothesized that children whose mothers have low education levels will score significantly lower than children whose mothers have high school or greater education levels. It is further hypothesized that AA children will score similar to their EUA peers and that differences are mostly due to maternal education levels.

METHOD

Participants

The study included two hundred and ten 4-year-old children attending a lottery-funded public school prekindergarten program in one urban county and in one rural county in Northeast Georgia. Children ranged in age from 4;0 to 5;2 (M = 4;6, SD = 0;5). The sample was 50% male and 50% female and 57.6% AA and 42.4% EUA. Nine percent of the children in the total sample had diagnosed special needs, per teacher report, and were receiving services through the schools for speech, attention deficit hyperactivity disorder (ADHD), developmental delays, or physical disabilities. Of the students with special needs, 37% were AA and 63% were EUA. Thirty percent of the total sample received free or reduced lunch. Of the students receiving free or reduced lunch, 59% were AA and 41% were EUA. According to information presented by parents at prekindergarten registration, 10% of the total sample had mothers with less than a high school education. Of these children, 67% were AA and 33% were EUA. Fifty-seven percent of children in the total sample had mothers who completed high school or a GED. Of these children, 63% were AA and 37% were EUA. Eleven percent of children had mothers who had some college or technical training. Of these children, 54% were AA and 46% were EUA. Finally, 22% of the total sample had mothers who completed college or graduate degrees. Of these children, 41% were AA and 59% were EUA. All children were reported to use English as their first language.

Measures

Children were administered the PPVT–III, Form A (Dunn & Dunn, 1997), and the EVT (Williams, 1997) within the first 45 days of the school year. According to the test manual, in the 4- to-5-year-old age range, the PPVT–III and the EVT report median internal reliabilities of .95 and .93, respectively, and a median correlation between the two instruments of .76. The

PPVT–III and EVT manuals indicate that the tests were standardized to have a mean score of 100 (SD = 15). The racial and ethnic representation in the standardization sample of both measures consisted of 18.1% AA, 12.9% Hispanic, 64.4% EUA, and 4.6% other. In the standardization sample, 17.1% of parents had less than a 12th-grade education, 31.3% had a high school diploma or GED, 31.3% had 3 years of college or technical education, and 20.3% had 4 or more years of college. Maternal education level was used to represent SES.

The demographic data used in the study were obtained from self-reports provided by parents during prekindergarten registration and by teachers regarding special needs. A report of maternal education levels and ethnicity was also obtained from the registration information. Information regarding whether the child received free or reduced lunch was obtained from school records.

Procedures

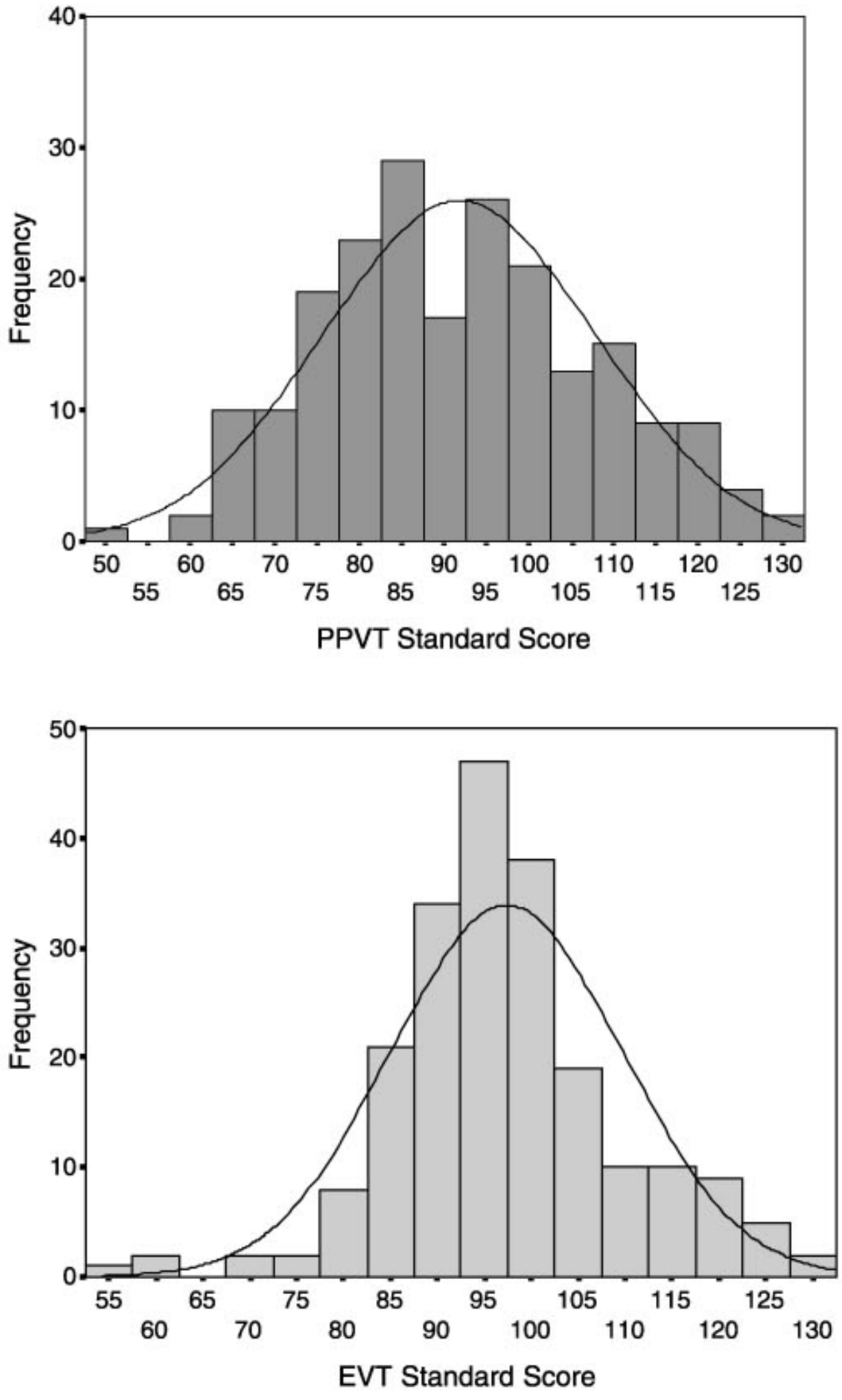
The PPVT–III and EVT were administered according to directions in the test manuals and norming procedures; accordingly, the PPVT–III was administered before the EVT in the same testing session, and therefore the tests were not counterbalanced. Graduate and undergraduate students with experience working with preschoolers were trained to administer the tests in two 4-hr training sessions. They were observed by the authors in testing adults, older children, and then preschool children before testing children from the study, and while testing the children from the study during the first 2 weeks of testing. To ensure administration and scoring accuracy, all test protocols were scored by one of the authors and then were re-checked by a graduate student who was trained in scoring. Raw scores were converted to standard scores using the procedures described in the test manual. Standard scores were calculated by using the accompanying computer software packages for the PPVT–III and EVT. Standard scores were calculated based on the child's chronological age at the time of testing.

Assenting children were taken from their classroom and administered the test in a quiet area at their school. The PPVT–III/EVT combination was part of a larger battery of preliteracy assessments that was administered to the children. The PPVT–III and EVT were administered early in the battery. The PPVT–III and EVT each required approximately 15–20 min to administer. Children were provided with stickers upon completing each test and with children's books for their participation in the study.

RESULTS

The relationship between standard scores on the PPVT–III and EVT and demographic variables was examined for each test separately. Figure 1 displays the distribution of scores for the PPVT–III and EVT. Table 1 shows the means and standard deviations for the standard scores received by children on the PPVT–III and EVT as a function of ethnicity, maternal education level, and gender. Figure 2 plots these results as a function of ethnicity and maternal education level for the PPVT–III. A 2 × 4 × 2, Ethnicity (AA vs. EUA) × Maternal Education Level (< high school, high school/GED degree, some college or technical school, college or graduate degree) × Gender (male vs. female) analysis of variance (ANOVA) was conducted for standard scores

Figure 1. Distribution of standard scores for the Peabody Picture Vocabulary Test—III (PPVT–III; Dunn & Dunn, 1997) and the Expressive Vocabulary Test (EVT; Williams, 1997).



obtained on the PPVT–III. This analysis yielded a significant model, $F(7, 190) = 7.46$, $p < .0001$, $\eta_p^2 = .22$. (For all η_p^2 effects reported, effects of 1% are considered small, effects of 9% or greater are medium, and effects of 25% or greater are large; Cohen, 1988). There was a nonsignificant main effect for gender and nonsignificant interactions with other factors, all $F(1, 194) < 1$, $p > .10$, $\eta_p^2 = .003$. There was a moderate-sized 18-point main effect for child ethnicity, $F(1, 187) = 43.510$, $p < .001$, $\eta_p^2 = .183$, with EUA children scoring at a mean of 102 (SD = 15) and AA children scoring at a mean of 84 (SD = 13). There was also a moderate main effect for maternal education level, $F(3,194) = 10.51$, $p < .001$, $\eta_p^2 = .14$.[1]

Post hoc Tukey tests comparing differences between means indicated that children whose mothers did not complete high

[1] Gender was not significant in any analysis conducted for the PPVT-III. Henceforth, our findings were collapsed across gender.

Table 1. Children's mean and standard deviation standard scores on the Peabody Picture Vocabulary Test—III (PPVT–III; Dunn & Dunn, 1997) and the Expressive Vocabulary Test (EVT; Williams, 1997) as a function of ethnicity, maternal education level, and gender.

	Assessment			
	PPVT–III		*EVT*	
Group	M	SD	M	SD
Ethnicity				
African American	84.21	12.79	93.83	11.78
European American	101.84	11.49	102.18	11.49
Maternal education level				
Less than high school	77.95	13.02	90.57	7.72
High school/GED	89.54	13.64	95.46	11.23
Technical	91.63	18.00	97.17	14.83
College	103.52	15.18	105.50	11.79
Gender				
Male	90.95	16.42	95.55	11.85
Female	92.40	15.83	99.15	12.61

school scored lower than children whose mothers completed high school, $t = 3.90$, $p = .001$; some college or technical school, $t = 3.65$, $p = .002$; or college, $t = 7.74$, $p < .001$. Children whose mothers completed college scored higher than children whose mothers completed some college or technical school, $t = 3.77$, $p = .001$; completed high school, $t = 6.42$, $p < .001$; or did not complete high school, $t = 7.74$, $p < .001$. In fact, the difference between the average score of the children of mothers who did not complete high school and those of mothers who graduated from college was 26 points. The interaction between ethnicity and maternal education level was nonsignificant, $F(3, 194) < 1$, $p > .10$. Thus, differences were found between groups according to ethnicity and to maternal education level on the PPVT–III,

Figure 2. Mean PPVT–III standard scores by ethnicity and maternal education level.

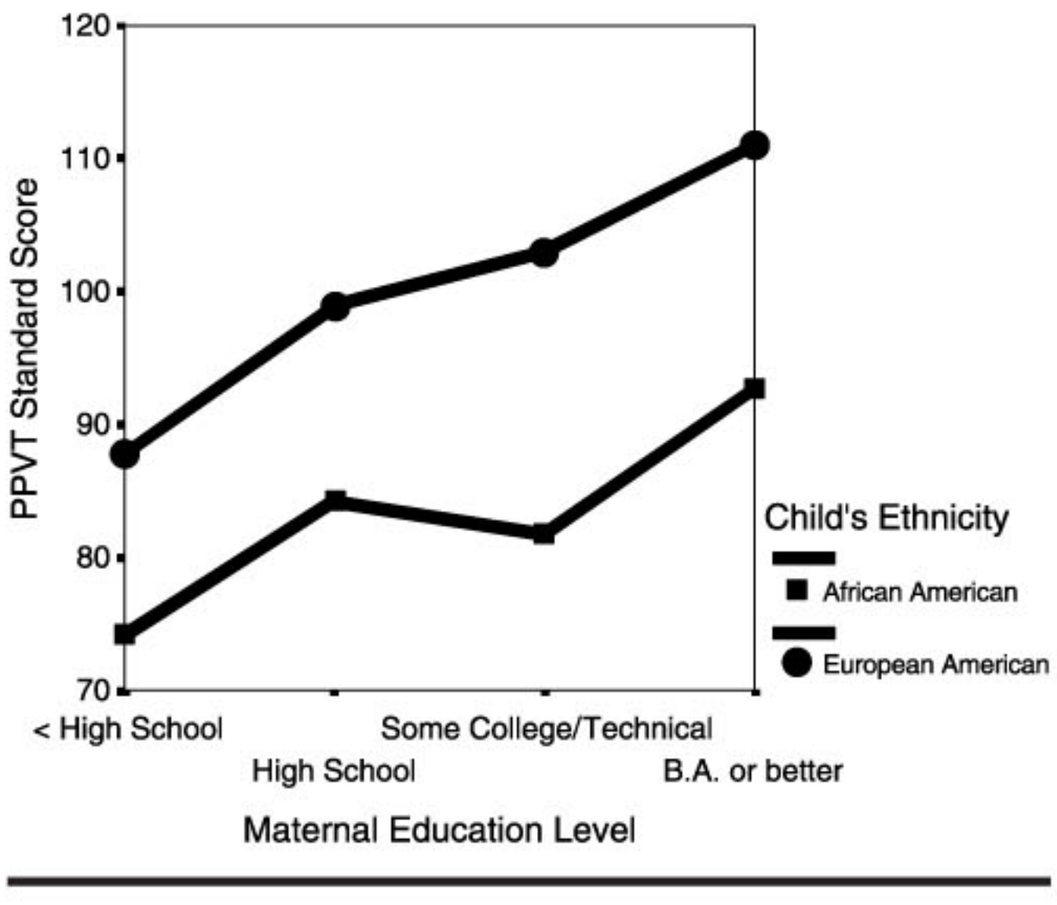


Table 2. Percentage of children displaying the presence of a significant vocabulary problem.

Group	*PPVT–III*	*EVT*
Ethnicity		
African American	24	5
European American	4	0
Maternal education level		
Less than high school	6	1
High school	18	3
Some college/technical	3	1
Bachelor's degree or better	1	0

suggesting possible bias against AA children and increasing levels of bias as a function of decreasing maternal education levels.

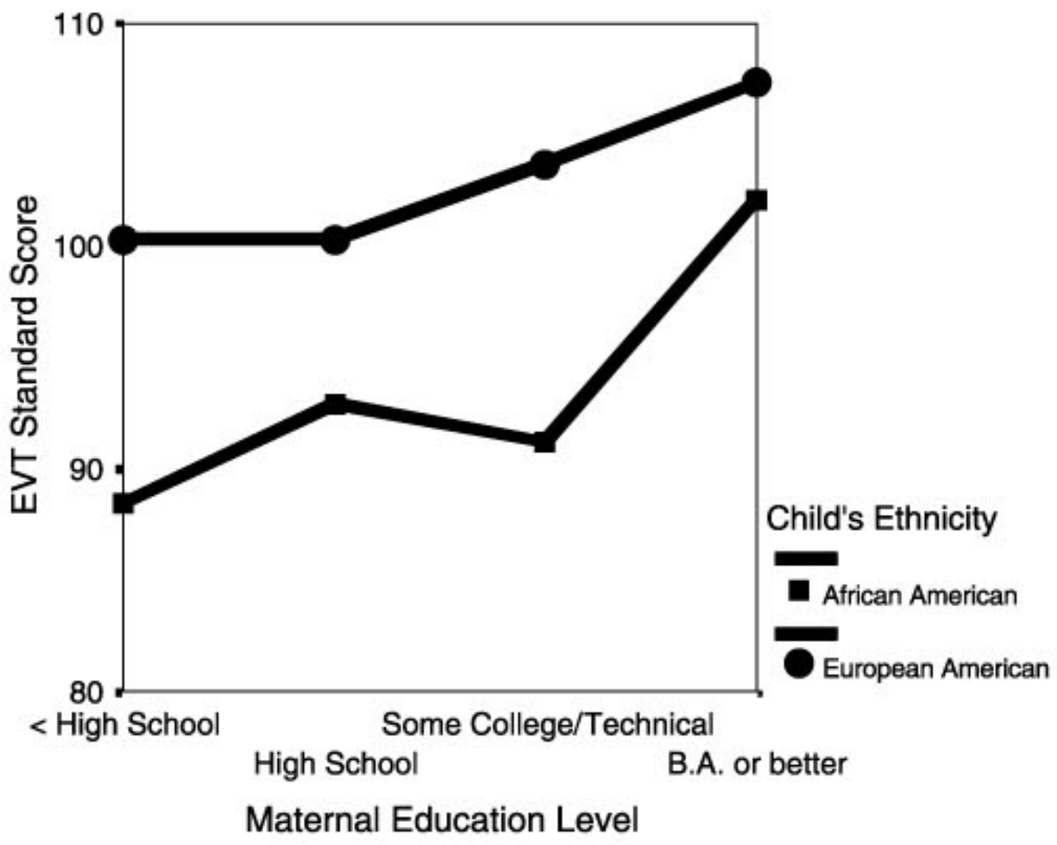
Another indicator of potential bias is the disproportional presence of very low scores on an assessment for a particular subgroup of children when a particular assessment is used, indicating the likelihood of a clinical problem with vocabulary. Ethnic differences appeared when using the PPVT–III to infer the clinical presence of vocabulary difficulties. Using a standard score of 80 or lower as an indicator of low vocabulary skills (the cutoff score often recommended by schools to indicate concern), 28.1% of our sample could be identified as having low vocabulary skills. However, as shown in Tables 2 and 3, most children displaying significant vocabulary problems were AA, $\chi^2(1) = 27.91$, $p \leq .001$, and most were children of mothers having less education, $\chi^2(3) = 23.123$, $p \leq .001$.

Similar analyses were conducted for the EVT. The means and standard deviations by ethnic group and maternal education level are depicted in Figure 3 and Table 1. A 2 × 2 × 4 (Gender × Ethnicity Group × Maternal Education Level) ANOVA was conducted using the standard score of the EVT as the dependent variable. This analysis yielded a small significant 4-point main effect for gender, $F(1, 194) = 3.90$, $p = .05$, $\eta_p^2 = .02$, favoring girls. There was also a small significant main effect for child ethnicity, $F(1, 194) = 16.55$, $p < .001$, $\eta_p^2 = .079$, such that AA children scored 8 points lower than their EUA counterparts on average. AA children scored 93.83 (*SD* 11.78) and EUA children scored 102.18 (*SD* 11.49). There was a small main effect for

Table 3. Percentage of children displaying the presence of a significant discrepancy between standard scores of the PPVT–III and the EVT.

Group	*PPVT–III higher*	*Test similar*	*EVT higher*
Ethnicity			
African American	2	30	26
European American	9	25	9
Maternal education level			
Less than high school	1	3	6
High school	5	33	18
Some college/technical	1	6	4
Bachelor's degree or better	3	12	7

Figure 3. Mean EVT standard scores by ethnicity and maternal education level.



maternal education level as well, $F(3, 194) = 5.82$, $p = .001$, $\eta_p^2 = .083$.

Post hoc Tukey tests indicated that children whose mothers have a college degree had higher standard scores than children whose mothers have some college or technical education, $t = 2.98$, $p = .017$; those whose mothers have a high school degree, $t = 5.21$, $p < .001$; and those whose mothers did not complete high school, $t = 5.11$, $p < .001$. On average, children having mothers with college degrees scored 15 points higher than children whose mothers did not graduate from high school. None of the interactions between ethnicity and maternal education level were significant, all $F < 1$, $p > .10$. We found significant effects of ethnicity and maternal education level on children's standard scores, but they were smaller than they were for the PPVT–III.

Using the EVT for inferring the clinical presence of low vocabulary skills seems to pose a smaller problem than using the PPVT–III. Using a standard score of 80 or lower as an indicator of low vocabulary, only 5.2% of our sample would be identified as having a significant vocabulary problem. As shown in Table 2, all of these children were AA, $\chi^2(1) = 8.54$, $p = .003$. However, maternal education level was not a significant factor in being identified as having low vocabulary, $\chi^2(3) = 3.88$, $p = .274$.

One striking finding is that children tended to score lower on the PPVT–III than on the EVT, $F(1, 209) = 39.18$, $p < .001$, $\eta_p^2 = .158$. However, the PPVT–III and EVT were co-normed such that children should have similar scores on each. One would not expect to find that the receptive test would be consistently more difficult than the expressive test (e.g., Gray et al., 1999). Consequently, to examine this issue more closely, we calculated discrepancy scores between the PPVT–III and the EVT for each child by subtracting the PPVT–III from the EVT. The EVT manual indicates that we can consider an 11-point difference between the standard scores of the EVT and the PPVT–III a *significant discrepancy* ($p < .05$) at 4;0, a 12-point difference at 4;5, and a 15-point difference at 5;0. Using these age-appropriate discrepancy scores, we found that 34.8% of our sample scored significantly lower on the PPVT–III compared to the EVT. By contrast, only 11% of our sample scored significantly lower on the EVT than on the PPVT–III.

When these discrepancy scores were examined, there were more AA children with significant discrepancies as defined by the EVT manual in favor of the EVT than EUA children, $\chi^2(1) = 25.12$, $p < .001$. Moreover, a Test × Ethnicity ANOVA indicated a significant interaction between these two factors: $F(1, 208) = 28.91$, $p < .001$, $\eta_p^2 = .125$, a moderate size effect. AA children performed better on the EVT than they did on the PPVT–III, but EUA children obtained nearly identical scores on both instruments. When examined as a function of maternal education level, a Test × Maternal Education Level ANOVA indicated a significant interaction between the factors, $F(3, 206) = 3.28$, $p = .022$, $\eta_p^2 = .046$, such that the standard scores discrepancies between the two tests grew larger as maternal education levels declined. After adjusting alpha for the number of contrasts computed to .0125 (.05/4), Bonferroni tests comparing EVT with PPVT scores indicated that children whose mothers have only a high school degree performed better on the EVT than on the PPVT, $t(20) = 5.07$, $p < .001$, $\eta_p^2 = .537$, as did children whose mothers did not graduate from high school, $t(118) = 4.81$, $p < .001$, $\eta_p^2 = .179$. However, children whose mothers have some college/technical degree or a college degree performed similarly on both tests, $t(23) = 2.09$, $p = .048$, $\eta_p^2 = .160$ and $t(45) = .971$, $p = .337$, $\eta_p^2 = .021$, respectively. On the other hand, there was no tendency for *significant* discrepancies in children's scores as defined by the EVT manual to vary as a function of maternal education level, $\chi^2(6) = 7.18$, $p = .305$. Putting these analyses together, it can be concluded that the PPVT–III is a more difficult test than the EVT. It is particularly more difficult for AA children and, perhaps, for children whose mothers have only a high school education or less.

To test for the possibility that our test results were different from those of Washington and Craig (1999) because they excluded children in special education, we re-analyzed the data eliminating children who were receiving special education services. The scores continued to be significantly lower for AA children on the PPVT, $F(1, 191) = 49.34$, $p < .001$, $\eta_p^2 = .212$. Children whose mothers had less than a high school education had lower PPVT scores also, $F(3, 191) = 14.77$, $p < .001$, $\eta_p^2 = .195$. Unlike in the study by Washington and Craig, the interaction between maternal education level and ethnicity in our study was not significant, $F(3, 191) = 1.60$, $p = .191$, $\eta_p^2 = .026$.

Item Analysis

Given evidence described above that the PPVT–III is possibly more difficult for AA children than EUA children compared to the EVT and over maternal education levels, it was important to determine whether cultural bias existed in the form of item bias. During the test construction process, item analysis is used to identify poorly performing items (too easy, too difficult, non-discriminating, and/or biased); such items are either removed from the test or modified to ensure that they are not more difficult for the specific group in question. In the current study, the goal of conducting this item analysis is different in that our purpose was to discern whether there were items on the test that functioned as culturally biased items for the children in our study. If items reflecting bias can be identified, then test bias is suggested.

Several statistical procedures have been developed to attempt to deal with the issue of creating equal-ability groups for comparison purposes. One popular approach is the Mantel-Haenszel (M-H) procedure (Holland & Thayer, 1986; Mantel & Haenszel, 1959). In this χ^2 approach, students are stratified by ability and the proportion correct for each item is compared across groups within each stratum. In most cases, the variable used for stratification is the total score on the test in question. This assumes, of course, that not all of the items are biased (i.e., that the test is a valid measure of the latent trait being measured).

In practice, the M-H typically divides the sample into quintiles and compares the focal and reference group performances on each item within each quintile with a 2×2 χ^2 table. The five tables' statistics are aggregated to produce the final M-H statistic, which is used to flag items suspected of DIF.

An item response theory alternative to the M-H has been suggested by Linacre and Wright (1989). They showed that the Rasch non-iterative normal approximation algorithm (PROX; Wright & Stone, 1979), by using all of the information from all students in both groups, provides an estimate for the difficulty of each item for each group and the standard error of the difference. The resulting statistic can be compared to the *t* distribution.

The PPVT–III data were analyzed using WINSTEPS (Linacre, 2002), a popular Rasch analysis program. Although the highest performing student in the sample reached item 96, in order to maintain an *N* of at least 25 in each group, only items 13 (the 4-year-old start item) through 84 were examined. In an initial analysis, we checked to see that the students and items fit the Rasch model well. This was accomplished by examining the INFIT (information-weighted) and OUTFIT (outlier-influenced) mean squares for test items and students. Linacre (2002) suggests 0.5 to 1.5 as an acceptable range. All items except one fell within this range. However, 20 students had OUTFIT mean squares greater than 1.5. These students were removed and a second analysis was run. The item parameters from the second analysis were used for the DIF analysis. In the second analysis, all items had acceptable fit statistics.

To interpret DIF analyses, Linacre (2002) recommends being as conservative as possible in interpreting the significance of the *t* statistic and suggests dividing the *t* statistic by 1.12 (sqrt(1.25)). Using an alpha of .05, ten items showed significant DIF. Three of these favored (i.e., were less difficult for) EUA students; seven favored AA students. Thus, the item analyses suggest that the difficulty of the PPVT–III for AA children in the current study could not be attributed to the overwhelming presence of culturally biased items.

DISCUSSION

Test Bias

By ethnic group. The results of the current study provide mixed evidence with regard to the issue of ethnic bias for the PPVT–III and EVT against AA children in the Southeast. Although the results of the study did not find evidence of ethnic bias against AA children in terms of item analysis, of particular concern was the AA children's performance on the PPVT–III. The performance of AA and EUA children was significantly different across most maternal education levels. These results contrast with those of Washington and Craig's (1999) PPVT–III study that found that only AA children from homes with mothers having less than a high school education scored significantly below the norm. The mean score of the AA children from our sample was almost identical to that found for a group of Southeastern kindergarteners by Campbell et al. (2001). However, neither Washington and Craig nor Campbell et al. carried out differential item functioning analyses, so it is unclear whether the differences they found were attributable to item biases or other adverse impact factors. Our findings indicated that, although the mean scores of AA children were considerably lower on the PPVT–III than might be predicted based on normative samples, these low scores could not be attributed to item bias. Therefore, other sources of bias may need to be investigated. Further, it is possible that in Campbell et al.'s study, the source of difference came from other risk factors such as SES or maternal education level rather than ethnic differences.

Several possibilities can account for the differences between studies. Although the age group that Washington and Craig (1999) used was very similar to that of our sample, there were differences between groups concerning the inclusion of children receiving special education. To test for this possibility, we re-analyzed the data eliminating children receiving special education services, and the scores on the PPVT–III continued to be significantly lower for AA children. This ethnicity effect was not larger for children having mothers with lower education levels.

A second possible explanation may be that the normative sample did not have adequate representation of children from the Southeastern United States. However, the manuals for the PPVT–III and EVT sample indicated that AA children had been adequately represented in the normative sample (see also Campbell, 1998). In addition, the results of the current study are consistent with those from other parts of the country (Campbell et al., 2001; Champion et al., 2003), indicating that the cause of the ethnic differences is not inadequate sample representation of children in the Southeast. However, because AA children only made up 18% of the normative sample, it was important to rule out the possibility that item selection may have had a negative impact on the minority sample scores (Reynolds, 2000). Our DIF analysis found no such biased items. There were more items favoring AA children (7) than EUA children (3).

Although there was a significant difference between the performance of AA and EUA children on the EVT, their performance was within 1 *SD* from the normative mean, except for children whose mothers had less than a high school education. These results are interesting in light of the fact that the two assessments were co-normed. In our study, the EVT appeared to be an easier test for AA children than the PPVT–III. It should also be pointed out that dialect differences between tester and child are less likely to have as much of an impact on expressive vocabulary because, once the general task is understood, the children themselves must produce a word that will be readily understood by the more linguistically sophisticated adult listener.

These results are consistent with those of studies of other language measures that continue to find differences against AA children, such as the Preschool Language Scale—3 (Qi et al., 2003), The Preschool Language Assessment Instrument (Fagundes et al., 1998), and the Test of Language Development—Primary (TOLD–P; Hammer et al., 2002). In fact, across subtests of the

TOLD–P, Hammer et al. found that the Picture Vocabulary and Oral Vocabulary subtests adversely affected AA children the most in their study of 6-year-old children. Taken together, these results indicate that vocabulary, particularly in the receptive modality, may be more vulnerable to cultural/linguistic differences in the AA population, despite documented grammatical differences in dialects between nonstandard dialects in AA children and a more "standard dialect" group.

By parent education. It is well documented that children whose mothers have low education levels tend to have lower vocabularies (Hart & Risley, 1995) and tend to score lower on standardized receptive and expressive vocabulary measures than children whose mothers have higher education levels (Campbell et al., 2002; Champion et al., 2003; Hammer et al., 2002; Qi et al., 2003; Roberts et al., 1999; Washington & Craig, 1999; Whitehurst, 1997). Maternal education is believed to be more stable and more strongly related to language development than income. This effect is consistent across racial groups, ethnic backgrounds, and rural or urban settings. Despite the consistency of the finding with previous research, these results indicate that using vocabulary measures to evaluate children's language for the purpose of identifying language disorders or to obtain estimates of verbal ability may lead to inaccurate conclusions or biased interpretations. Moreover, those children who are AA and who live in poverty are therefore at a greater risk of being misdiagnosed or being diagnosed with language disorders when they do not have one. Further, the validity of using vocabulary tests for this purpose is questionable because they are not sensitive to language disorders (Gray et al., 1999; Ukrainetz & Bloomquist, 2002).

By gender. The results of this study indicate that the PPVT–III is not biased for gender, although the EVT favors girls slightly. These results are consistent with previous research that favors girls in verbal abilities (e.g., Qi et al., 2003). Unfortunately, the test developers did not report a test for gender differences, and therefore, it is unclear whether our results are consistent with those of the normative sample. In contrast, Gray et al. (1999) found no gender bias on the EVT but found gender differences on the PPVT–III (the direction of the difference is not reported), indicating that this factor may be quite variable across samples, although when there are differences, they tend to favor girls.

Test discrepancies by parent education and ethnic background. A surprising finding in our sample was that 34% of children in the sample scored significantly higher on the EVT than on the PPVT–III, based on the norms in the test manual and the cutoff score of 80. This trend was particularly noticeable among AA children and somewhat among children whose mothers had lower education levels. These results were unexpected because the typical pattern of performance in these measures is either that of no difference between assessments (because the tests were co-normed) or toward receptive vocabulary scores being greater than expressive vocabulary scores (Gray et al., 1999; Ukrainetz & Bloomquist, 2002). Williams (1997), the EVT test author, noted that persons who have higher EVT scores than PPVT–III scores may be better at demonstrating knowledge in an open rather than a focused format (p. 40). It is reasonable to suggest that the open format of the EVT might be more conducive for AA children and children whose parents have lower education levels.

Although some researchers have documented discrepancy between the PPVT–III and the EVT, the number of children showing significant discrepancies between the two tests typically is small and is in favor of the PPVT–III, at least among children who are developing typically. Gray et al. (1999), for example, found that 1% of the children with typical language scored higher on the EVT than on the PPVT–III. However, 32% of the children with specific language impairment demonstrated a significant EVT > PPVT–III discrepancy. Similarly, Ukrainetz and Bloomquist (2002) found that 5 of 27 children scored higher on the EVT than on the Receptive One-Word Picture Vocabulary Test (Gardner, 2000), and only 1 child of 27 had the EVT > PPVT–III profile. Our findings suggest that such discrepancies in favor of the EVT might be higher for AA children and children whose mothers have less than a high school education.

Although we might be tempted to assume that AA children and children whose mothers have low education levels have smaller vocabularies, and Hart and Risley (1995) have documented that children from lower socioeconomic levels do in fact tend to hear less vocabulary throughout the day than children who grow up in higher socioeconomic levels, we must consider the cultural assumptions that the above measures entail. It is possible that different socialization practices in the families lead to limited experiences with the types of tasks used in the tests, which may have had differential adverse impact on the scores. Therefore, we must develop a deeper understanding of how these socialization practices can dramatically affect performance on what would seem to be similar kinds of tests, although different in modality of assessment. The receptive test is seemingly more difficult than the expressive test for our AA children, who often respond better to more contextualized types of questions rather than decontextualized tasks such as the PPVT–III.

Purcell-Gates (1996), for example, indicated that low-SES parents tended to do more shared reading and explicit literacy instruction only after children entered elementary school and received direct literacy instruction. These results indicate possible differences between low-SES preschool children's ability to respond to more decontextualized tests that require them to listen to an adult and to point to pictures than higher SES children, who may have these experiences before formal schooling. Purcell-Gates also found that children from low-SES parents tended to participate in literacy experiences as part of the contextual focus of oral discourse rather than as a separate activity in which literacy, listening, and pointing to pictures is the whole focus of the activity (Purcell-Gates, 1996).

CLINICAL IMPLICATIONS

The current findings lead us to strongly caution practitioners in the use of the PPVT–III for verbal ability estimates or screening and for identification of language disorders (e.g., Campbell, 2002; Champion et al., 2003; Gray et al., 1999; Ukrainetz & Bloomquist, 2002) when assessing AA children and children whose mothers have less than a high school education. These concerns are reinforced when other reports indicate that the PPVT–III may overestimate EUA children's vocabulary (Ukrainetz & Bloomquist, 2002; Ukrainetz & Duncan, 2000).

Clinicians may want to use alternative forms of assessment such as analyzing the number of different words used in conversation and narratives (Ukrainetz & Bloomquist, 2002); examining the use of noncontrastive grammatic search forms

(Seymour, Bland-Stewart, & Green, 1998); using dynamic assessment (Fagundes et al., 1998; Peña, Iglesias, & Lidz, 2001); and using measures that are specifically developed for the fair assessment of children, including AA children, such as the Diagnostic Evaluation of Language Variation (Seymour, Roeper, & de Villiers, 2003), or perhaps the EVT.

Although the test authors have indicated that significant differences between the PPVT–III and the EVT may indicate a retrieval problem when the EVT is lower than the PPVT–III, discrepancies in the opposite direction may indicate differences in how the child demonstrates knowledge. Therefore, in terms of determining whether an AA child or a child with low maternal education has significant vocabulary difficulties, we suggest that EVT scores may be more representative of children's true vocabulary level.

Finally, it is recommended that test developers provide norms for children whose mothers have less than a high school education, and provide more robust norms for different racial and ethnic groups, given that despite the use of national census ratios, we still find differences across groups that may be due to the underrepresentation of minorities in the samples (Reynolds, 2000). In addition, investigators and test developers need to determine whether dialect mismatch between tester and child impacts performance to the extent that we saw in our study; if so, the implications for testing in our schools are serious. If this is the case, the norms should also account for these differences.

CONCLUSION

In summary, the results of our study indicate that AA children and children whose mothers have less than a high school education tend to score low on the PPVT–III and on the co-normed measure, the EVT. However, because scores on the EVT tend to place such children within the typical range and within 1 *SD* of the sample mean, we prefer it over the PPVT–III as an indicator of a child's "true vocabulary" skill. According to our findings, the PPVT–III has too great a tendency to place AA children and children whose mothers have low education levels at risk for being unfairly identified as presenting with a potential language disorder. Therefore, practitioners should use alternative assessment methods such as nonformal and dynamic assessment or use great caution interpreting these test results when evaluating these children. The EVT might serve as one alternative.

ACKNOWLEDGMENT

Funding for this project was provided by the U.S. Department of Education Early Childhood Educator Professional Development Program, Award S349A010167.

REFERENCES

Anderson-Yokel, J., & Haynes, W. O. (1994). Joint book-reading strategies in working class African American and White mother–toddler dyads. *Journal of Speech and Hearing Research, 37,* 583–593.

Blank, M., Rose, S., & Berlin, L. (1978). *Preschool Language Assessment Instrument*. New York: The Psychological Corporation.

Brooks-Gunn, J., Klebanov, P. K., & Duncan, G. J. (1996). Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics. *Child Development, 67,* 396–408.

Brown, R., Reynolds, C. R., & Whitaker, J. S. (1999). Bias in mental health testing. *School Psychology Quarterly, 14*(3), 208–238.

Campbell, J. M. (1998). Review of the Peabody Picture Vocabulary Test, Third Edition. *Journal of Psychoeducational Assessment, 16,* 334–338.

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*(1), 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,* 121–127.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test, Third Edition*. Circle Pines, MN: American Guidance Service.

Fagundes, D. D., Haynes, W. O., Haak, N. J., & Moran, M. J. (1998). Task variability effects on the language test performance of Southern lower socioeconomic class African American and Caucasian 5-year-olds. *Language, Speech, and Hearing Services in Schools, 29,* 148–157.

Gardner, M. F. (2000). *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications.

Goff, G. S., & Montague, J. C., Jr. (1980). Caution in using the PPVT (Form A) with rural preschool Black children. *Journal of Educational Research, 74,* 23–26.

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.

Hammer, C. S., Penncock-Roman, M., Rzasa, S., & Tomblin, J. B. (2002). An analysis of the Test of Language Development—Primary for item bias. *American Journal of Speech-Language Pathology, 11,* 274–284.

Hart, B., & Risely, T. (1995). *Meaningful differences in everyday parenting and intellectual development in young American children*. Baltimore: Brookes.

Heath, S. (1983). *Ways with words*. Cambridge, MA: Cambridge University Press.

Hickman, J. A., & Reynolds, C. R. (1987). Are race differences in mental test scores an artifact of psychometric methods? A test of Harrington's experimental model. *Journal of Special Education, 20,* 409–430.

Holland, P., & Thayer, D. (1986, April). *Differential item performance and the Mantel-Haenszel procedure.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Huston, A. C., McLoyd, V. C., & Garcia Coll, C. T. (1994). Children and poverty: Issues in contemporary research. *Child Development, 65,* 275–282.

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed., pp. 145–166). Needham Heights, MA: Allyn & Bacon.

Kaufman, A. S., & Kaufman, N. L. (1983). *Brief form manual for the Kaufman Assessment Battery for Children*. Circle Pines, MN: American Guidance Service.

Linacre, J. M. (2002). WINSTEPS Rasch-Model Computer Program [Computer software]. Chicago, IL: MESA.

Linacre, J. M., & Wright, B. D. (1989). Mantel-Haenszel DIF and PROX are equivalent! *Rasch Measurement Transactions, 3*(2), 52–53.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719–748.

Mercer, J. R. (1979). *The system of multicultural pluralistic assessment: Conceptual and technical manual*. New York: The Psychological Corporation.

Mount-Weitz, J. (1996). Vocabulary development and disorders in African American children. In A. Kamhi, K. E. Pollock, & J. L. Harris (Eds.), *Communication development and disorders in African American children* (pp. 189–226). Baltimore: Brookes.

National Center for Education Statistics. (1999). *The early childhood longitudinal study*. Retrieved from http://www.nces.ed.gov/ecls

Osterlind, S. J. (1983). *Test item bias*. Newbury Park, CA: Sage.

Peña, L., Iglesias, A., & Lidz, C. S. (2001). Reducing test bias through dynamic assessment of children's word learning ability. *American Journal of Speech-Language Pathology, 10,* 138–154.

Purcell-Gates, V. (1996). Stories, coupons, and the TV guide: Relationships between home literacy experiences and emergent literacy knowledge. *Reading Research Quarterly, 31,* 406–428.

Qi, C. H., Kaiser, A. P., Milan, S. E., Yzquierdo, Z., & Hancock, T. B. (2003). The performance of low-income, African American children on the Preschool Language Scale—3. *Journal of Speech and Hearing Research, 46,* 576–590.

Reynolds, C. (2000). Methods for detecting and evaluating cultural bias in neuropsychological tests. In E. Fletcher-Jansen, T. Strickland, & C. R. Reynolds (Eds.), *Handbook of crosscultural neuropsychology* (pp. 317–334). New York: Kluwer Academic/Plenum Publishers.

Reynolds, C. R., Lowe, P. A., & Saenz, A. L. (1999). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (3rd ed., pp. 549–595). New York: John Wiley & Sons.

Roberts, J. E., Bruchinal, M., & Durham, M. (1999). Parent's report of vocabulary and grammatical development of African-American preschoolers: Child and environment associations. *Child Development, 70,* 92–106.

Seymour, H. N., Bland-Stewart, L., & Green, L. J. (1998). Difference versus deficit in child African American English. *Language, Speech, and Hearing Services in Schools, 29,* 96–108.

Seymour, H. N., Roeper, D., & de Villiers, J. (2003). *Diagnostic Evaluation of Language Variation*. San Antonio, TX: The Psychological Corporation.

Skiba, R. S., Knesting, K., & Bush, L. (2002). Cultural competent assessment: More than nonbiased tests. *Journal of Child and Family Studies, 11*(1), 61–78.

Stockman, I. J. (2000). The new Peabody Picture Vocabulary Test–III: An illusion of unbiased assessment? *Language, Speech, and Hearing Services in Schools, 31,* 340–353.

Taylor, O., & Lee, D. L. (1987). Standardized tests and African-American children: Communication and language issues. *The Negro Educational Review, 38,* 67–80.

Ukrainetz, T. A., & Bloomquist, C. (2002). The criterion validity of four vocabulary tests compared with a language sample. *Child Language Teaching and Therapy, 18,* 59–79.

Ukrainetz, T. A., & Duncan, D. S. (2000). From old to new: Examining score increases on the Peabody Picture Vocabulary Test–III. *Language, Speech, and Hearing Services in Schools, 31,* 336–339.

U.S. Department of Education. (2004a). *Early reading first.* Retrieved December 3, 2004, from http://www.ed.gov/programs/earlyreading/index.html

U.S. Department of Education. (2004b). *Migrant education even start.* Retrieved December 3, 2004, from http://www.ed.gov/programs/mees/index.html

U.S. Department of Health and Human Services. (2003). *Head Start FACES 2000: A whole-child perspective on program performance. Report for the Administration for Children and Families.* Retrieved December 3, 2004, from http://www.acf.hhs.gov/programs/core/ongoing_research/faces/faces004thprogress/

Washington, J. A. (2001). Early literacy skills in African-American children: Research considerations. *Learning Disabilities Research & Practice, 16,* 213–221.

Washington, J. A., & Craig, H. K. (1992). Performance of low-income, African American preschool and kindergarten children on the Peabody Picture Vocabulary Test—Revised. *Language, Speech, and Hearing Services in Schools, 23,* 75–82.

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

Whitehurst, G. J. (1997). Language processes in context: Language learning in children reared in poverty. In L. B. Adamson & M. A. Romski (Ed.), *Communication and language acquisition: Discoveries from atypical development* (pp. 233–266). Baltimore: Brookes.

Williams, K. T. (1997). *Expressive Vocabulary Test*. Circle Pines, MN: American Guidance Service.

Williams, K. T., & Wang, J. (1997). *Technical references to the Peabody Picture Vocabulary Test—Third Edition*. Circle Pines, MN: American Guidance Service.

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: MESA Press.

Received February 10, 2004
Accepted February 25, 2005
DOI: 10.1044/0161-1461(2006/003)

Contact author: Dr. María Adelaida Restrepo, Department of Speech and Hearing Science, Arizona State University, P.O. Box 870102, Tempe, AZ 85287. E-mail: laida.restrepo@asu.edu