# CONCURRENT VALIDITY OF THE PEABODY PICTURE VOCABULARY TEST–THIRD EDITION AS AN INTELLIGENCE AND ACHIEVEMENT SCREENER FOR LOW SES AFRICAN AMERICAN CHILDREN

Jonathan M. Campbell
Stephen K. Bell
The University of Memphis

Lori K. Keith
University of Tennessee, Memphis

Authors tested the validity of the Peabody Picture Vocabulary Test–Third Edition (PPVT-III) as a screening measure for intelligence and achievement. The PPVT-III and Kaufman Assessment Battery for Children (KABC) were administered to 416 African American children of low socioeconomic status in a counterbalanced design. Results indicated that the PPVT-III correlated .58 with the KABC Mental Processing Composite (MPC) score; however, participants scored significantly lower ($M$ = 8.3 points) on the PPVT-III than on the MPC. Although receiver operating characteristic (ROC) analyses supported the use of the PPVT-III as a valid intellectual and achievement screener, the selection of a single cutoff score for the PPVT-III was problematic. The purpose of the screening program should guide selection of a cutoff score for the PPVT-III.

*Keywords:* Peabody Picture Vocabulary Test–Third Edition (PPVT-III), validity, screener, minority, children, Kaufman Assessment Battery for Children (KABC)

Screening is defined as the identification of unrecognized problems through the use of brief measures in order to distinguish between those persons who probably suffer from a disorder from those who probably do not (Derogatis & DellaPietra, 1994). The purpose of screening is to identify those persons who are at greater risk for a problem, not to yield a diagnosis (Satz & Fletcher, 1988). The practice of screening "well" populations is built on the notion that early problem detection is advantageous by leading to early intervention and improved outcomes. Based on the premise that early educational interventions produce positive outcomes, early screening of children for educational

Jonathan M. Campbell, Department of Psychology, The University of Memphis (now at the Yale Child Study Center); Stephen K. Bell, Department of Psychology, The University of Memphis; Lori K. Keith, University of Tennessee, Memphis.

This work was conducted with the support of The University of Memphis Van Vleet Memorial Doctoral Fellowship. Additional support was provided by grants from the Bureau of Maternal and Child Health (USPHS-MCJ-479158-04), the Administration on Developmental Disabilities (OHDS Grant-90-DD-0364-04), the Robert Wood Johnson Foundation, the Smith Richardson Foundation, and the Hearst Family Foundation.

Correspondence concerning this article and requests for offprints should be addressed to Jonathan M. Campbell, Yale Child Study Center, 230 South Frontage Road, P.O. Box 207900, New Haven, CT 06520-7900.

problems has been endorsed as a service to children who have special needs (Lichtenstein & Ireton, 1991).

Although educational screening instruments can vary in terms of coverage, desirable measures are those that are reliable, valid, standardized, and cost effective (Lichtenstein & Ireton, 1991). For screening measures, the most important type of psychometric validity is that of predictive validity, the ability to identify those with the disorder of interest and exclude those who do not have the disorder. Cost effectiveness is also important in screening. Useful screening measures are those that cost little in terms of materials, administration time, and level of expertise required for administration and scoring (Lichtenstein & Ireton, 1991; Satz & Fletcher, 1988).

The most recent revision of the Peabody scales, the Peabody Picture Vocabulary Test–Third Edition (PPVT-III; Dunn & Dunn, 1997), appears to possess several of these desirable characteristics. Like its predecessors, the PPVT-III is a standardized measure that is quick, inexpensive, and can be administered and scored by nonprofessionals (Dunn & Dunn, 1997). In addition, test authors specify the following two purposes for the PPVT-III: (a) as a brief achievement measure of receptive, or hearing vocabulary; and, (b) as a screening measure of verbal ability (Dunn & Dunn, 1997).

Numerous studies endorsed the validity of the Peabody Picture Vocabulary Test–Revised (PPVT-R; Dunn & Dunn, 1981) as a screening measure of intelligence and achievement (e.g., Williams & Wang, 1997). The PPVT-R was found to correlate positively and significantly with varied measures of intelligence (*r*s ranged from .23 to .78; Dunn & Dunn, 1997) and achievement (*r*s ranged from .33 to .80; Williams & Wang, 1997). Despite positive validity findings, however, the PPVT-R tended to yield lower scores for minority children. For example, Washington and Craig (1992) found that low income African American preschool and kindergarten children performed more than one standard deviation below the mean of the PPVT-R standardization sample and concluded that the PPVT-R was inappropriate for use with this population. During field testing of the PPVT-III, test authors attempted to reduce racial disparity by eliminating items that were found to be biased against racial and ethnic groups (Dunn & Dunn, 1997).

The first purpose of the present study is to investigate the concurrent validity of the PPVT-III by determining the extent to which it compares with measures of intelligence and achievement. The second purpose is to examine how the PPVT-III compares with prior versions in measuring intelligence and achievement in low socioeconomic (SES) minority children. The third purpose is to analyze the accuracy of the PPVT-III as a screening measure of intellectual ability and scholastic achievement in children.

## Method

### Participants

Participants were 213 boys (51.2%) and 203 girls (48.8%) from the Memphis, Tennessee area who had recently completed kindergarten. Participants ranged in age from 70 to 89 months ($M$ = 75.86 months; $SD$ = 3.12 months) and were selected for study inclusion if their mothers identified themselves as African American. Participants' median household income from all sources was $10,000 per year ($M$ = $13,467; $SD$ = $13,400), and the majority of participants' parents were receiving governmental assistance at the time of testing (e.g., Aid for Dependent Children, AFDC). At the time of testing, 34% of children's mothers had not earned a high school degree, 42% of children's mothers completed high school or earned a General Equivalency Diploma, and 23.8% had completed one or more years of education beyond high school.

Children's mothers were participants in a larger study that examined the efficacy of a nurse-delivered, home intervention designed to reduce negative health-related outcomes such as childhood injuries. Mothers were selected for study inclusion if they met at least two sociodemographic risk factors (i.e., unmarried; < 12 years of education; unemployed; see Kitzman et al., 1997). Data were collected as part of a follow-up assessment to determine the efficacy of the nurse-visitation program. As data collection coincided with children's natural transition from kindergarten to structured first-grade education, the assessment matched the timing when school systems are likely to conduct intellectual and achievement screenings with students.

### Procedure

Advanced graduate students administered a standard battery of tests to all participants as part of the follow-up assessment. The third author, a doctoral level, licensed school psychologist, trained graduate students to administer and score the PPVT-III and the Kaufman Assessment Battery for Children (KABC; Kaufman & Kaufman, 1983). The PPVT-III and KABC were individually administered in counterbalanced order.

Examiners completed four videotaped, pilot test administrations that were reviewed for standardized administration and scoring by the third author prior to data collection. To further insure standardized data collection, all test administrations were videotaped and randomly reviewed by the third author. All test protocols were rechecked by an independent reviewer to insure scoring accuracy. The third author resolved scoring discrepancies via a third review of the protocol.

### Measures

*PPVT-III.* The PPVT-III is an individually administered, norm-referenced test designed for use with persons ages 2½ to 90 years. The PPVT-III includes two parallel forms, Form III-A and Form III-B, each containing 204 test items. Form III-A was administered to all participants. The respondent is required to select one of four pictures that best represents the word spoken by the examiner. The PPVT-III yields standard scores (i.e., $M = 100$; $SD = 15$), takes approximately 10 to 15 minutes to complete, and is easy to administer and score (Campbell, 1998; Dunn & Dunn, 1997). The PPVT-III demonstrates sound internal consistency reliability as evidenced by a median alpha coefficient of $r = .95$ across age groups and good temporal stability as evidenced by a median test-retest coefficient of $r = .92$ across age groups (Campbell, 1998). Authors of the PPVT-III also provide sound evidence for content, construct, internal, and criterion-related validity (Campbell, 1998).

*KABC.* The KABC is an individually administered test of intelligence and achievement (Kaufman & Kaufman, 1983). The KABC yields a Mental Processing Composite (MPC) IQ score that is further divided into Sequential Processing (SEQ) and Simultaneous Processing (SIM) factor scores. The KABC also yields an Achievement Composite score comprised of the following subtests at age 6 years: Faces and Places, Arithmetic (MATH), Reading/ Decoding (READ), and Riddles (RIDD). The MPC, SEQ, SIM, MATH, READ, and RIDD scores were calculated for all participants. All scores are standardized ($M = 100$; $SD = 15$). The psychometric properties of the KABC are sound and well documented. For example, mean internal consistency reliability coefficients exceed .91 for the MPC and .92 for the READ subtest (e.g., Kaufman & Kaufman, 1983; Lichtenberger & Kaufman, 2000). Also, multiple concurrent validity studies have documented strong relationships between the KABC and measures of intelligence (e.g., Kaufman & Kaufman, 1983; Lichtenberger & Kaufman, 2000).

### Statistical Analyses

To examine first-order relationships between measures, Pearson product-moment correlations between PPVT-III and KABC scores were calculated. A series of paired-sample $t$ tests, with Bonferroni correction, contrasted PPVT-III and KABC MPC, MATH, READ, and RIDD mean scores.

*Receiver Operating Characteristic (ROC) analyses.* ROC analyses were conducted to test the diagnostic accuracy of the PPVT-III as a screen for (a) intellectual ability deficits, using the KABC MPC as criterion and (b) scholastic achievement deficits, using KABC achievement subtests as criteria. ROC analysis was developed from statistical decision theory for use with radar signal detection experiments (Swets, 1992) and has been applied to examine the accuracy of various biomedical technologies (e.g., Swets, 1988), self-report measures of psychopathology (e.g., Somoza, Steer, Beck, & Clark, 1994), and psychoeducational diagnostic indicators (e.g., Watkins, Kush, & Glutting, 1997). ROC analysis graphically represents paired true-positive (i.e., sensitivity) and false-positive ratios across a test's full range of decision thresholds (i.e., all possible cutoff points). A measure with random accuracy produces a line that bisects the ROC graph, referred to as a random ROC or line of no information (Hsiao, Bartko, & Potter, 1989). As diagnostic accuracy improves, the ROC curve deviates from the random ROC line moving toward the upper left-hand corner of the graph. A test that perfectly discriminates between two diagnostic groups yields an ROC curve that connects points (0,0), (0,1), and (1,1) on the graph.

Several objective indices exist to quantify the accuracy of an ROC curve (e.g., Macmillan & Creelman, 1991), the most widely used index involves calculating the area under the ROC curve (AUC; Swets, 1992). AUCs can range in value from .50 (random accuracy) to 1.0 (perfect accuracy) and can be interpreted intuitively as the probability of a screening test score correctly classifying a pair of participants, one "diseased" and one normal (Colliver, Vu, & Barrows, 1992). AUCs ranging from .5 to .7 denote low test accuracy, those ranging from .7 to .9 denote moderate test accuracy, and those .9 to 1.0 denote high test accuracy (Swets, 1988; Watkins et al., 1997).

The *Diagnostic and Statistical Manual of Mental Disorders–Fourth Edition* (*DSM–IV*; American Psychiatric Association, 1994) was consulted to determine cutoff scores for KABC measures. The *DSM–IV* defines significant subaverage intellectual functioning as an IQ of about 70 or below on one or more standardized, individually administered intelligence tests such as the KABC. *DSM–IV* guidelines were used to define an intellectual deficit hit as corresponding to a KABC MPC score less than or equal to 70; an achievement deficit hit corresponded to a KABC MATH, READ, or RIDD score less than or equal to 70. Sensitivity and false-positive ratios were calculated for PPVT-III standard and raw scores obtained by at least one participant in the sample. Fifty-nine data points were present in the standard score ROC and 70 data points were used in the raw score ROC. The authors then estimated AUCs using the trapezoidal method described by Hsiao et al. (1989); standard errors for AUCs were calculated according to the method proposed by Hanley and McNeil (1982) and used by Weinstein, Berwick, Goldman, Murphy, and Barsky (1989).

## Results

Pearson product-moment correlations between the PPVT-III and KABC MPC and scholastic achievement measures were moderately positive and significant, ranging from .44 to .64 (see Table 1). The correlations parallel those observed between the PPVT-III and measures of cognitive ability and oral language ranging from .43 to .83 (Dunn & Dunn, 1997). Pearson correlation coefficients were also similar to those observed between the PPVT-R and KABC MPC (*M r* = .58; Williams & Wang, 1997) and KABC achievement subtest scores (e.g., *r*s ranged from .41 to .62; Bing & Bing, 1985; Mcloughlin & Ellison, 1984).

Table 1
*Correlation Coefficients Between PPVT-III and KABC Measures*

| | PPVT-III | MPC | SEQ | SIM |
|---|---|---|---|---|
| PPVT-III | – | | | |
| MPC | .58 | – | | |
| SEQ | .44 | .82 | – | |
| SIM | .55 | .90 | .49 | – |
| | PPVT-III | RIDD | MATH | READ |
| PPVT-III | – | | | |
| RIDD | .64 | – | | |
| MATH | .55 | .46 | – | |
| READ | .48 | .41 | .72 | – |

*Note.* PPVT-III = Peabody Picture Vocabulary Test, Third Edition; KABC = Kaufman Assessment Battery for Children; MPC = Mental Processing Composite; SEQ = Sequential Processing; SIM = Simultaneous Processing; RIDD = Riddles Achievement Subtest; MATH = Arithmetic Achievement Subtest; READ = Reading/Decoding Achievement Subtest. All Pearson's correlations are two-tailed, *df* range from 411 to 413 due to missing data, $p < .001$.

Similar to the PPVT-R (e.g., Bing & Bing, 1985; Sattler, Hilson, & Covin, 1985; Washington & Craig, 1992), low SES, African American children scored more than one standard deviation below the mean of the PPVT-III standardization sample, *M* = 82.26 (*SD* = 12.19). As with the PPVT-R, participants scored significantly lower on the PPVT-III when contrasted with their KABC MPC, MATH, and READ performances (Bing & Bing, 1985; see Table 2).

Table 2
*Means, Standard Deviations, and t values Denoting Difference Between PPVT-III and KABC Measures*

| Measure | *M* | *SD* | Mean diff | *t* value |
|---|---|---|---|---|
| PPVT-III | 82.26 | 12.19 | – | – |
| MPC | 90.58 | 11.30 | -8.32 | -15.39** |
| RIDD | 83.45 | 8.85 | -1.19 | -2.57* |
| MATH | 88.85 | 12.74 | -6.59 | -11.27** |
| READ | 94.10 | 12.52 | -11.84 | -19.04** |

*Note.* PPVT-III = Peabody Picture Vocabulary Test–Third Edition; KABC = Kaufman Assessment Battery for Children; MPC = Mental Processing Composite; RIDD = Riddles Achievement Subtest; MATH = Arithmetic Achievement Subtest; READ = Reading/Decoding Achievement Subtest; DIFF = difference. All *t* tests are from paired-sample contrasts with *df* = 411, two-tailed with Bonferroni correction for multiple comparisons.
*$p < .05$. **$p < .01$.

As an intellectual screener, the PPVT-III yielded an AUC of .909 with a standard error of .051 using standard scores (see Figure 1); as an achievement screener, the PPVT-III yielded an AUC of .823 with a standard error of .031 using standard scores (see Figure 2). Therefore, as an intellectual screener, the probability with which the PPVT-III correctly classified a pair of children (i.e., one scoring > 70 on MPC, one scoring ≤ 70 on MPC) was 91%. As an achievement screener, the probability that the PPVT-III correctly classified a pair of children (i.e., one above and one below 70 on at least one KABC achievement subtest) was 82%. Hanley and McNeil's (1983) method to compare AUCs was used, and results did not differ when raw scores were used for intellectual screening, AUC = .907, *SE* = .052, *z* = 0.08, *ns*, or achievement

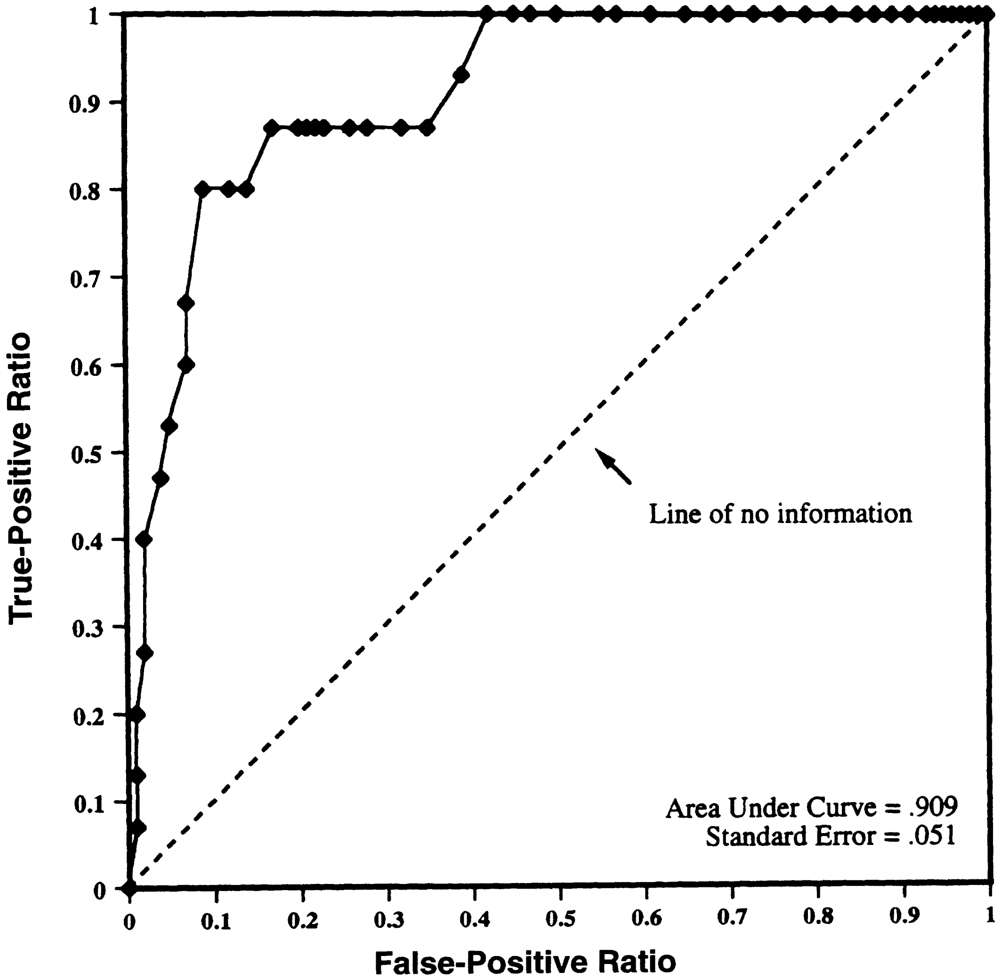

*Figure 1.* Receiver operating characteristic (ROC) curve using Peabody Picture Vocabulary Test–Third Edition (PPVT-III) standard scores for intellectual screening.

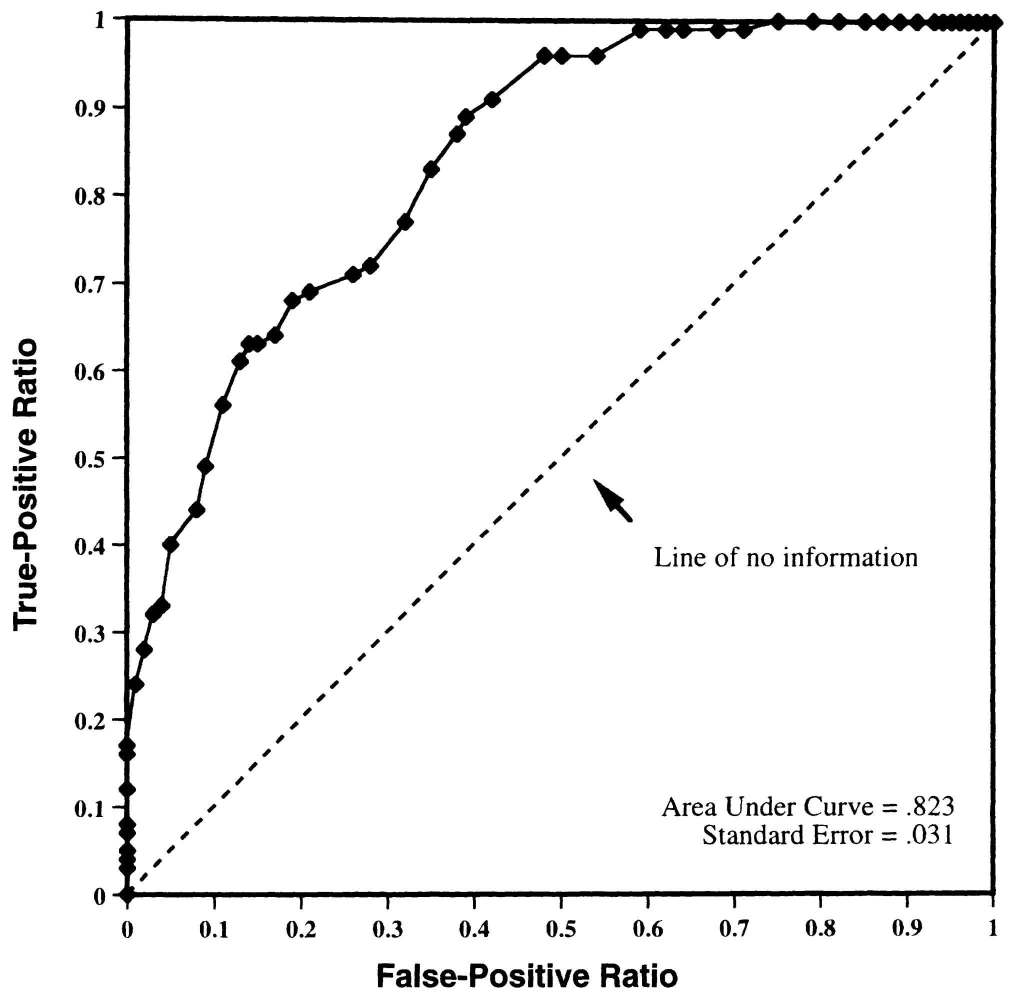

*Figure 2.* Receiver operating characteristic (ROC) curve using Peabody Picture Vocabulary Test–Third Edition (PPVT-III) standard scores for achievement screening.

screening, AUC = .843, *SE* = .029, *z* = 1.28, *ns*. As an intelligence screener, the PPVT-III AUC compares with those obtained by diagnostic imaging techniques (e.g., .90–.98), such as computed tomography or mammography, used to discriminate diseased from normal tissue (Swets, 1988). The achievement AUC compares with those obtained by weather forecasts, which range from .75 to .90 for predicting rain, and .65 to .80 for predicting temperature (Swets, 1988). Due to the similar results produced by both standard and raw score cutoffs and the use of standard scores for making screening decisions in applied settings, the remainder of the results are reported for standard scores only.

Although the PPVT-III produced moderate to high AUC values, determining an optimal cutoff score was problematic. Carran and Scott (1992) suggest that indicators of sensitivity, specificity, positive predictive value (PPV), and overall hit rate approach .80 for a screening test to be considered valuable. As an intelligence screener, sensitivity, specificity, and hit rate values met or exceeded the .80 criterion at scores 67 through 71; however, the highest PPV among these scores was .25 (12/48; see Table 3). The highest PPV of .60 (3/5) was observed at score 51, resulting in excellent specificity and overall hit rate but low sensitivity. As an achievement screener, the PPVT-III did not achieve a pair of sensitivity and specificity values that met or exceeded .80. As an achievement screener, the highest PPV of 1.00 (9/9) was observed at score 60, resulting in perfect specificity and good overall hit rate but unacceptable sensitivity (see Table 4). No PPVT-III cutoff score met all criteria suggested by Carran and Scott (1992) as an intelligence or achievement screener.

Table 3
*Accuracy of Selected Cutoff Scores When Screening for Intellectual Deficits*

| Cutoff score | Sensitivity | Specificity | PPV | NPV | Hit rate |
|---|---|---|---|---|---|
| 51[a] | .20 | .99 | .60 | .97 | .96 |
| 67[b,c] | .80 | .91 | .25 | .99 | .91 |
| 68[b] | .80 | .88 | .20 | .99 | .88 |
| 69[b] | .80 | .86 | .17 | .99 | .86 |
| 70[b] | .87 | .83 | .16 | .99 | .83 |
| 71[b] | .87 | .80 | .14 | .99 | .80 |

*Note.* PPV = positive predictive validity; NPV = negative predictive validity. Base rate of MPC scores at or below 70 = 3.6%.
[a]Cutoff score that yielded highest overall hit rate. [b]Cutoff score that achieved sensitivity, specificity, and hit rate at or above .80 per Carran and Scott's (1992) guidelines. [c]Cutoff score that yielded referral rate 3 times higher than base rate per Lichtenstein and Ireton's (1991) guidelines.

Table 4
*Accuracy of Selected Cutoff Scores When Screening for Achievement Deficits*

| Cutoff score | Sensitivity | Specificity | PPV | NPV | Hit rate |
|---|---|---|---|---|---|
| 60[a] | .12 | 1.00 | 1.00 | .84 | .84 |
| 63[b] | .24 | .99 | .78 | .85 | .85 |
| 75[c] | .68 | .81 | .45 | .92 | .79 |

*Note.* PPV = positive predictive validity; NPV = negative predictive validity. Base rate of one or more achievement scores at or below 70 = 18.1%. [a]Cutoff score that yielded highest PPV. [b]Cutoff score that yielded highest overall hit rate. [c]Cutoff score that yielded referral rate 1.5 times higher than base rate per Lichtenstein and Ireton's (1991) guidelines. No cutoff score achieved sensitivity, specificity, and hit rate at or above .80 per Carran and Scott's (1992) guidelines.

## Discussion

Generally, findings indicate that the PPVT-III is comparable to the PPVT-R as evidenced by similar relationships with KABC summary scores and like performance by low income, African American children. Despite the attempts to reduce racial differences, the PPVT-III appears to perform similarly to prior editions of the Peabody scales. On average, the PPVT-III tends to underestimate both intellectual ability and scholastic achievement, as measured by the KABC, in low SES, African American children. Although the PPVT-III produced acceptable AUCs, the test did not meet established criteria at any single cutoff score.

It is important to note, however, that most screening instruments do not meet or exceed criteria established by Carran and Scott (1992). For example, the Brigance K&1 Screen (Brigance, 1992) achieved adequate specificity, .82, and overall hit rate, .80, values but low sensitivity, .67, and PPV, .41, when used to detect special education status in preschoolers (Mantzicopoulos, 1999). Two reviews point to similar problems with other early childhood screening instruments, such as the Denver Development Screening Test and Early Screening Inventory (Carran & Scott, 1992; Gredler, 1997). On average, early childhood screening instruments achieve sensitivity values in the .48 to .63 range, specificity values in the .89 to .91 range, and PPVs from .42 to .65.

The PPV of screening instruments, including the PPVT-III, is impacted negatively when detecting low base-rate disorders. The PPV of a screening test falls sharply as the prevalence of the disorder of interest decreases. Even when screening tests demonstrate excellent sensitivity and specificity (e.g., .95), low prevalence of a disorder necessarily limits its PPV. For example, a screening test with sensitivity and specificity of .95 yields a PPV of .50 when screening for a disorder with prevalence rate of 5% (Derogatis & DellaPietra, 1994). That is, 50% of those identified at-risk by the screen will not be diagnosed with the disorder. When more realistic sensitivity, .80, and specificity, .90, values are used to illustrate the problem of low base rates, PPVs fall to approximately .30 (Clark & Harrington, 1999; Derogatis & DellaPietra, 1994). In the present study, 3.6% of the sample scored below 70 on the MPC, and 18.1% scored below 70 on one or more of the achievement subtests; therefore, the poor performance of the PPVT-III at any single cutoff score is due partly to inherent problems of detecting low base-rate disorders.

Before discussing strategies for dealing with the problem of detecting low base-rate disorders, it is important to establish different costs associated with two types of screening errors, false negatives and false positives. False negative errors result in children missing needed special education services. Given the importance of detecting educational problems early, some propose that false negatives are the more serious of the two errors and should be minimized when screening (Lichtenstein & Ireton, 1991; Rafoth, 1997). Proponents argue that the primary purpose of a screening program is to identify those children who may benefit from special services and assert that a false positive error is correctable through follow-up evaluation, whereas a false negative error cannot be corrected. False positive errors create problems as well. In terms of financial resources and manpower, false positives strain follow-up assessment efforts and can restrict resources available for special education services. Ironically, a screening program striving to detect a high percentage of children with a disorder could result in reduced direct services available for special education. In addition to strained resources, Gredler (1997) states that false positive errors can cause undue parental stress, anger, and worry associated with unnecessary evaluation, whereas Lichtenstein and Ireton (1991) suggest that a false positive result can produce a negative "self-fulfilling prophecy" that influences follow-up assessments.

When the prevalence rate for a problem is low, one strategy for reducing overall misclassification errors is to operate at a cutoff score that compromises sensitivity. On the ROC, these scores are located on the lower left-hand corner of the curve (Harber, 1982; Metz, 1978; Swets, 1992). Stringent cutoffs restrict the number of persons misclassified as false positives, thereby reducing costs associated with undue parental stress and unnecessary follow-up testing such as administration of an intelligence test. Using this strategy with our sample, a cutoff of 51 as a screen for intelligence yielded the highest overall classification rate of .96 with corresponding specificity of .99, PPV of .60, and a sensitivity of .20 (see Table 3). A cutoff score of 63 as a screen for achievement yielded the highest overall classification rate of .85 with specificity of .99, PPV of .78, and a sensitivity of .24 (see Table 4). The stringent cutoff strategy effectively limits problem detection by assuming that misclassification errors are equally undesirable. In applied settings it is unlikely that misclassification errors will be assigned equal importance; therefore, those who use the PPVT-III as a screener may wish to establish cutoff scores according to other guidelines.

When screening for low base-rate disorders, Lichtenstein and Ireton (1991) suggested operating at a cutoff score that yields referral rates 1.5 to 2.5 higher for disorders with base rates ranging from 5% to 10%. For base rates as low as 1% to 3%, Lichtenstein and Ireton suggest using a cutoff score that yields referral rates as high as 3 times the base rate. Compared with the stringent cutoff strategy, Lichtenstein and Ireton's guidelines are designed to minimize false-negative errors while increasing costs related to the evaluation of false positives. In our sample, cutoff scores were selected to yield referral rates 3 times higher for intellectual deficits and 1.5 times higher than the base rate for achievement deficits (see Tables 3 and 4). As expected, the increase in referrals yielded a significant increase in test sensitivity and false-positive errors. If detection of low intellectual functioning is considered paramount regardless of the corresponding increase in costs associated with evaluating false positives, one might reasonably use a cutoff score of 67. In this case, using a cutoff score of 67 results in a referral rate three times greater than the base rate and yields acceptable sensitivity, specificity, and hit-rate values. In our sample, a cutoff of 67 produced a false-positive ratio of 75% in those referred but correctly identified 80% of those children scoring at or below 70.

Study conclusions should be tempered due to several limitations. First, the sample consisted of 6-year-old, low SES, African American children thereby limiting generalizations to this population. Children of low SES status tend to show deficits in vocabulary skills (e.g., Hart & Risley, 1999), a finding that may account for the significant differences between scores on the PPVT-III and KABC regardless of minority status. It is not possible, however, to attribute the results of the study to either low SES status or minority status as these two variables were confounded in the sample.

Second, the KABC may not be the best measure for use as a criterion measure for the sample. It was normed over a decade earlier than the PPVT-III, a difference between measures that may account for a portion of the mean differences observed in the study. Flynn (1984) documented that norms increase in difficulty over time at a rate of approximately 3 standard score points per decade, the so-called "Flynn effect." Therefore, mean differences observed between measures may result, in part, because the PPVT-III is a newer test than the KABC as opposed to being a significantly different test. It is also important to note that the KABC measures intellectual functioning and achievement with minimal verbal requirement. Because the KABC does not sample a wide range of vocabulary skills, the PPVT-III may perform better as a screening instrument if the criterion measure includes a greater sampling of vocabulary skills.

Also, a single test score is not equivalent to a diagnostic category. Participants' hypothetical follow-up status was determined by performance on one measure with no certainty that low scorers would satisfy all diagnostic criteria for a disorder. If one is interested in the diagnostic accuracy of the PPVT-III in detecting Mental Retardation, for example, KABC results satisfy only one diagnostic criterion for this disorder. In addition, our data do not address the utility of the PPVT-III as a screen for giftedness as no participant in this sample achieved an MPC score greater than 118. Finally, the present study examines the concurrent validity of the PPVT-III, one stage in validation of a screening instrument. As Satz and Fletcher (1988) noted, the most important type of validity for a screening instrument is predictive validity. According to Satz and Fletcher, predictive validity for a screening instrument should be assessed in a longitudinal prediction design with a follow-up interval of at least 3 years.

Given the large number of studies that investigated the validity of the PPVT-R (Williams & Wang, 1997) and its frequent selection as a screening instrument (May & Kundert, 1992), the PPVT-III is likely to be studied in some detail in the near future. Investigations designed to further assess the usefulness of the PPVT-III as a screening measure may address the limitations of the present study in several ways. First, future studies are needed to cross-validate the present findings by sampling a wider scope of participants. Studies that expand the range of participants' age, SES, and racial and ethnic diversity can test the generalizability of our findings. Second, future investigations may further test the diagnostic accuracy of the PPVT-III by administering the test in the context of a thorough assessment with the purpose of arriving at definitive diagnoses of intellectual or learning disabilities. Lastly, a longitudinal design should be used to examine the predictive validity of the PPVT-III in detecting subsequent intellectual delays, academic deficits, or both.

## References

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: Author.

Bing, S. B., & Bing, J. R. (1985). Comparison of the K-ABC and PPVT-R with Head Start children. *Psychology in the Schools, 22,* 245-249.

Brigance, A. (1992). *Brigance K & 1 Screen* (3rd ed.). North Billerica, MA: Curriculum Associates.

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

Carran, D. T., & Scott, K. G. (1992). Risk assessment in preschool children: Research implications for the early detection of educational handicaps. *Topics in Early Childhood Special Education, 12,* 196-211.

Clark, A., & Harrington, R. (1999). On diagnosing rare disorders rarely: Appropriate use of screening instruments. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 40,* 287-290.

Colliver, J. A., Vu, N. V., & Barrows, H. S. (1992). Screening test length for sequential testing with a standardized-patient examination: A receiver operating characteristic (ROC) analysis. *Academic Medicine, 67,* 592-595.

Derogatis, L. R., & DellaPietra, L. (1994). Psychological tests in screening for psychiatric disorder. In M. E. Maruish (Ed.), *The use of psychological testing for treatment planning and outcome assessment* (pp. 22-54). Hillsdale, NJ: Erlbaum.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test–Revised.* Circle Pines, MN: American Guidance Service.

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

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29-51.

Gredler, G. R. (1997). Issues in early childhood screening and assessment. *Psychology in the Schools, 34,* 99-106.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29-36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology, 148,* 839-843.

Harber, J. R. (1981). Assessing the quality of decision making in special education. *The Journal of Special Education, 15,* 77-90.

Hart, B., & Risley, T. R. (1999). *The social world of children learning to talk.* Baltimore: Brookes.

Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). Diagnosing diagnoses: Receiver operating characteristic methods and psychiatry. *Archives of General Psychiatry, 46,* 664-667.

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

Kitzman, H., Olds, D. L., Henderson, C. R., Hanks, C., Cole, R., Tatelbaum, R., McConnochie, K. M., Sidora, K., Luckey, D. W., Shaver, D., Engelhardt, K., James, D., & Barnard, K. (1997). Effect of prenatal and infancy home visitation by nurses on pregnancy outcomes, childhood injuries, and repeated childbearing. *Journal of the American Medical Association, 278,* 644-652.

Lichtenberger, E. O., & Kaufman, A. S. (2000). The assessment of preschool children with the Kaufman Assessment Battery for Children. In B. A. Bracken (Ed.), *The psychoeducational assessment of preschool children* (3rd ed., pp. 103-123). Needham Heights, MA: Allyn & Bacon.

Lichtenstein, R., & Ireton, H. (1991). Preschool screening for developmental and educational problems. In B. A. Bracken (Ed.), *Psychoeducational assessment of preschool children* (pp. 486-513). Boston: Allyn & Bacon.

Macmillan, N. A., & Creelman, C. D. (1991). *Detection theory: A user's guide.* New York: Cambridge University Press.

Mantzicopoulos, P. (1999). Reliability and validity estimates of the Brigance K&1 Screen based on a sample of disadvantaged preschoolers. *Psychology in the Schools, 36,* 11-19.

May, D. C., & Kundert, D. K. (1992). Kindergarten screenings in New York state: Tests, purposes, and recommendations. *Psychology in the Schools, 29,* 35-41.

Mcloughlin, C. S., & Ellison, C. L. (1984). Comparison of scores for normal preschool children on the Peabody Picture Vocabulary Test–Revised and the achievement scales of the Kaufman Assessment Battery for Children. *Psychological Reports, 55,* 107-114.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine, 8,* 283-298.

Rafoth, M. A. (1997). Guidelines for developing screening programs. *Psychology in the Schools, 34,* 129-137.

Sattler, J. M., Hilson, D. E., & Covin, T. M. (1985). Comparison of Slosson Intelligence Test–Revised norms and Peabody Picture Vocabulary Test–Revised with Black Headstart children. *Perceptual and Motor Skills, 60,* 705-706.

Satz, P., & Fletcher, J. M. (1988). Early identification of learning disabled children: An old problem revisited. *Journal of Consulting and Clinical Psychology, 56,* 824-829.

Somoza, E., Steer, R. A., Beck, A. T., & Clark, D. A. (1994). Differentiating major depression and panic disorders by self-report and clinical rating scales: ROC analysis and information theory. *Behaviour Research and Therapy, 32,* 771-782.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285-1293.

Swets, J. A. (1992). The science of choosing the right decision threshold in high-stakes diagnostics. *American Psychologist, 47,* 522-532.

Washington, J. A., & Craig, H. K. (1992). Performances of low-income, African American preschool and kindergarten children on the Peabody Picture Vocabulary Test–Revised. *Language, Speech, and Hearing Services in the Schools, 23,* 329-333.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly, 12,* 235-248.

Weinstein, M. C., Berwick, D. M., Goldman, P. A., Murphy, J. M., & Barsky, A. J. (1989). A comparison of three psychiatric screening tests using receiver operating characteristic (ROC) analysis. *Medical Care, 27,* 593-607.

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