Cross Validating a New Preschool Screening Test

Marcia Strong Scott, Shihfen Tu, and Kathryn L. Fletcher
University of Miami

Abstract: A new preschool screening battery, consisting of five brief cognitive tasks, was presented to a sample of four- and five-year-old children with, and without mild learning problems. The sample included 177 normally achieving children and 21 children classified as either learning disabled (n = 7) or developmentally delayed (n = 14). Cross-sample test validation was demonstrated when 81% of both educational groups were correctly classified. The White/nonHispanic group achieved significantly higher screening scores than an omnibus minority group even when controlling for testing language. However, Hispanic children tested in English had a significantly higher screening score than Hispanic children tested in Spanish or both English and Spanish, and this impacted the race/ethnicity comparisons. Using percentage of exact matches, the interrater agreement was 80% or greater for all but one of the tasks. It was greater than 90% for all five tasks when the criterion for an agreement between raters was achieving scores within 1 point of each other.

Although screening tests were commonly used in medical practice to “...detect the persons possessing certain characteristics as a first step for further examinations and action on these ‘positive’ cases” (Yerushalmy, 1947, p. 1433), they have also come to play an important role in the ascertainment of children at-risk for developmental disabilities and other related educational problems. As stated by Shepard (1997, p. 93), “Developmental screening can be thought of as a preliminary step in the identification of children as handicapped. If the initial, brief screening suggests that children may have serious learning or developmental problems, they are referred for a more in-depth developmental assessment.” This further evaluative procedure is critical for making a valid recommendation concerning the child’s subsequent educational trajectory, and is the only accepted consequence of the results of a screening test. Results of a screening test alone should never be, although they sometimes are, used to make decisions concerning, for example, labeling, diagnosis, inclusion/exclusion or placement (Gredler, 1997; Gridley, Mucha, & Hatfield, 1995; Rafter, 1997). Essentially, a developmental screening test attempts to identify that smaller subset of children in the general population for whom the time and expense of a more comprehensive educational evaluation is warranted.

Since the passage of Public Law 99-457 (1986), there has been increased motivation to develop valid screening instruments that would identify those preschool children who are eligible for, and who require, a specialized educational intervention. Such interventions provide economic benefits by treating potential problems before they mature (Sattler, 1992), thus minimizing their impact on subsequent academic performance. In particular, an emphasis has been on development of screening tests that would more effectively identify those children with only mild learning problems, the most difficult to detect with screening instruments (Mercer, Algozzine, & Trifiletti, 1988; Meisels, 1989). Included in this category would be those children classified as educable mentally retarded (EMH) or...
learning disabled (LD). The former group, of course, consists of children with a significant intellectual impairment (IQ scores between 2 and 3 SDs below the mean) that impacts a broad range of cognitive capabilities. The latter group, on the other hand, is composed of children with less severe intellectual impairments (IQs in the normal range) and with more limited and specific cognitive difficulties. More recently, due to changes in labeling practices, young children with labels of developmental delay (DD) have been added to the mildly impaired category for purposes of test evaluation.

Scott and her colleagues have been working on development of new screening tests that would more effectively identify these mildly impaired children before they suffer repeated failure in grade school and are placed in special education. Specifically, they are developing three screening tests; one appropriate for three-year-olds (Scott, Fletcher, & Martell, 2000; Scott & Fletcher, 2001), one for preschool four- and five-year-olds (e.g., Scott, Fletcher, Jean-Francois, Urbano, & Sanchez, 1998) and one for children in kindergarten (Scott, Fletcher, et al., 1998).

The assumption underlying the expected effectiveness of these new screening tests is that mild cognitive impairments will become detectable if the assessment device used requires the active engagement of the children’s cognitive processing capabilities (Greenfield & Scott, 1985). Based on this empirical premise, rather than any theoretical position, a wide variety of cognitive tasks were evaluated. Since children at risk for eventually falling into an EMH, SLD, or DD category were the specific target populations, all potential cognitive tasks were evaluated in terms of whether or not they contributed to the correct classification of children with one of these labels versus same age peers without any learning problems. Those tasks that did contribute to effective differentiation of the children were retained for further evaluation.

The recommended criterion for evaluating screening instruments is an evaluation of its effectiveness in classifying individual children into their correct classifications. Other methods of evaluation have significant drawbacks. For example, an evaluation of a screening test based on its correlation with a second instrument can be negatively impacted by “...the imperfect criterion to which the evaluated instrument is being compared,” (Bracken, 1987, p. 316). Furthermore, the correlation between a screening instrument and an achievement measure provides little information regarding a screening test’s ability to correctly identify which children are at-risk and which are not (e.g., Gredler, 1997; Lichtenstein, 1981; Limbos & Geva, 2001; Meisels, 1989; Satz & Fletcher, 1988). Therefore, the to-be-evaluated potential test items, like screening tests themselves, were evaluated in terms of their contribution to the accuracy with which children with, and without mild learning problems were classified.

The initial evaluations used children age six through eight. This slightly older group of children was selected because there was a large pool of potential test measures to evaluate and this required a large pool of children classified as either LD or EMH (e.g., Scott & Greenfield, 1991; Scott, Greenfield, Partridge, & Gold, 1991; Scott, Greenfield, & Sterental, 1986). These evaluations, and those that followed, were in terms of concurrent validity, a measure which has been described as being especially important in determining the value of a screening test (McIntosh, Gibney, Quinn, & Kundert, 2000).

Those tasks retained for further evaluation were grouped into three multi-task batteries and presented to preschool samples of children with and without designated learning problems (e.g., Scott, Deuel, Claussen, & Sanchez, 1993; Scott et al., 1996). Tasks that contributed to the classification accuracy of these younger samples of children were subsequently grouped into a single nine-task battery that constituted the first version of a new preschool screening test.

A subset of measures from this larger battery was associated with classification accuracy levels greater than 80% for the EMH and LD children and of greater than 90% for the normally achieving children in both the first (Scott, Deuel, Urbano, Fletcher, & Torres, 1998) and second (Deuel, 1997) version of the screening battery, demonstrating their
concurrent predictive validity. However, since a higher percentage of the children classified as EMH were correctly identified compared to those with an LD label, a third concurrent validity study was planned aimed at increasing sensitivity to children with LD. In this final concurrent study (Scott, Fletcher, et al., 1998), the cognitive measures in the battery were evaluated separately for preschool and kindergarten groups. The set of tasks associated with the best sensitivity (91%) and specificity (91%) for the preschool sample was different from the set of tasks that led to the highest level of sensitivity (87%) and specificity (77%) for the kindergarten sample. For both screening sets, the level of sensitivity was still higher for children with an EMH label than for those classified as LD.

The next step taken in the ongoing process of test development was an evaluation of the test-retest reliability of the two screening tests (Jean-Francois, Urbano, & Scott, 1999). A test-retest format was used to assess the consistency or stability of the tests over time (Carmines & Zeller, 1979) because an acceptable level of reliability is a necessary, if not sufficient condition for test validity. On all measures used to evaluate the two screening tests, the preschool screening instrument demonstrated acceptable levels of test-retest reliability while the kindergarten screening test did not.

Given these results, continued evaluations of the preschool screening test seemed warranted. Reported in this study is a cross-sample validation of the concurrent predictive validity of the five-task set selected in Scott, Fletcher, et al. (1998). Such a cross-sample validation is necessary because items were originally selected and evaluated based on characteristics of a particular sample. Another estimate of the sensitivity and specificity one can obtain with that set must be made with a new sample of children. The levels of classification accuracy achieved with the new sample will provide a good estimate of the levels of classification accuracy one could expect to achieve were the test actually administered to a preschool population where the children would be referred or not based on their performance.

Method

Participants

During the 1998-1999 school year, 198 children were administered the new five-component preschool screening test. Of these, 177 were attending one of eighteen private preschools scattered throughout Miami-Dade County. The other 21 participants were in special education classes in the Miami-Dade Public School System since free public education is available to, but only available to, preschool children with disabilities. These 21 children attended one of six public schools, were classified as either DD (14) or LD (7) and were drawn from the same population, as were the normally achieving preschool children.

In keeping with the idea that one should clearly define the children constituting the samples employed (McIntosh et al., 2000), the criteria for classifying a child as either LD or DD are provided.

For a child to be classified as LD, there must be a) a significant difference between language performance and other developmental levels, or b) a significant difference between receptive and expressive language abilities, or c) a significant language delay based on criteria presented in the test or evaluation manual.

For a child to be classified as DD, he or she must score 1 SD or more below the mean on two or more developmental areas, for example, language and cognition, motor performance and language. Performance on an IQ test would represent one measure of cognitive ability.

The children’s parent or guardian designated each of the 198 children participating as being White/Hispanic (n = 106), White/nonHispanic (n = 36), African American (n = 18), Haitian (n = 18), Multi-Racial (n =12), Black/Hispanic (n = 6) or Asian (n = 2). The 21 children in the exceptional sample were either African American (n = 2) or White/Hispanic (n = 19). The racial/ethnic distribution of the preschool children in this sample, most of whom were members of minority groups, was similar to the racial/ethnic distribution of children in Miami-Dade County Public School kindergarten classes where, in
the same school year, 54% were Hispanic, 32% were Black and only 14% were classified as White/Other in the public school database. The category “Other” included Asian, American Indian and Multiracial families.

Mean chronological age of the normally achieving sample at the time of testing was 56.5 months (range = 48-62; SD = 4.3) and of the exceptional sample was 55.8 (range = 48-62; SD = 3.4). One hundred and nine were male and 89 were female.

**Determining Testing Language**

For children in this sample, who were exposed to two languages, it was necessary to determine which language to use to administer the screening test. Decisions were based on the child’s ability to converse in the teacher nominated language, the child’s performance on a brief five-page picture task, and the examiner’s assessment of the child’s understanding of the screening test instructions. If a child did not appear to comprehend, the examiner switched to the child’s second language. If uncertainty was still present, examiners gave the instructions in both languages. Spanish was the only “other” language required. Examiners fluent in both Spanish and English used data sheets with instructions in Spanish. Of the 198 children tested, 173 were tested in English, 13 in Spanish and 12 in both English and Spanish. Regardless of the language used to administer the screening test, responses in either English or Spanish were accepted.

**Preschool Screening Test Contents, Instructions and Scoring**

The preschool screening test consisted of three generating and two identification tasks, all of which required the active engagement of the children’s cognitive processing skills. Four of the five tasks used colored photographs of meaningful pictures that appeared on 35.6 cm wide by 21.6 cm high laminated white paper pages. All pages were placed in a four-hole legal size black binder. The semantic information/verbal task was presented without any visual aids. Tasks are described below in the same order as they appeared in the screening battery.

*Taxonomic generation.* This first generating task consisted of just two pages. On the first were an ear and a foot and on the second, a skirt and a jacket. The examiner named the exemplars and then asked each child to name additional body parts and more clothing. In this and the other two generating tasks, the examiner probed twice to maximize children’s verbal responses. One point was awarded for each correct verbal response and 1/2 point for any nonverbal response where a child pointed to a particular body part or article of clothing.

*Picture pointing.* This task consisted of four pages, on each of which appeared six unrelated meaningful pictures that were placed in two horizontal rows of three pictures each. The children were asked to “Point to each picture just one time.” No points were awarded if there were any repetitions or omissions in the pointing sequence, one point was earned if the pointing was error free but not systematic, and 2 points were given if the pointing sequence was systematic. The dependent measure, pointing penalty, was the total number of points awarded over the four pages minus the total number of omissions and repetitions.

*Semantic information/verbal.* For this second generating task, the examiner said, “What does an airplane do?” Then, “What does an airplane look like?” Each correct verbal response was awarded one point. If in response to the first question the child made motions with his/her arms indicating flying, 1/2 point was awarded for that response. The dependent measure was the total number of points awarded.

*Picture Rhyming.* The picture-rhyming task consisted of one training page and two test pages. There were four pictures on each page; one pair was situated above, and one pair below a horizontal black line that bisected the page. The names of one of the picture pairs rhymed, while the names of the other pair did not. The examiner named the pictures twice as instructions were administered. Children were required to point to the picture pair whose names rhymed or sounded alike. For those tested in English, the English names of one pair rhymed and for those tested in Span-
ish, the Spanish names of one pair rhymed. One point was awarded for a correct choice on each of the two test pages.

Semantic information/visual. For the final generating task, the children were shown a page on which was a group of more than a dozen fruit. The children were first asked, “What can you do with fruit?” and then were asked, “What do fruits look like?” One point was awarded for each response that was correct or true of most fruit.

All children were tested individually, in a single session, by one of 14 female examiners. At the start of the session, children were shown a selection of award certificates and stickers that they could choose from after they completed the test. Administration time was about 10 minutes. During test presentation, children were periodically praised to both encourage and reward their participation. Testing continued throughout the entire 98-99 school year.

Results

Cross-Validating the Preschool Screening Test

The preschool screening test was cross validated by examining the level of classification accuracy achieved with the new and much larger sample of 177 normally achieving children and 21 exceptional children. To obtain accuracy measures, number of children in each of the two educational groups who achieved each score was displayed in a frequency distribution. Potential cut scores were then examined. In each case, all children whose scores fell below that score were called exceptional while all children with scores at or above the cut were called normally achieving. Classification accuracy was then determined by comparing the children’s classification based on their screening scores to their actual classification. The cut score selected was the one that resulted in the best combination of sensitivity (percentage of all exceptional children whose scores fell below the cut) and specificity (percentage of all normally achieving children whose scores fell at or above the cut). The aim, though, was to achieve the highest level of sensitivity that was consistent with an acceptable level of specificity.

When the cut score (23) that had been determined in the Scott, Fletcher, et al. (1998) study was applied to the distribution in this study, a sensitivity of 95% (20 of 21) was observed but the specificity was only 36% (63 of 177). This is perhaps not unexpected as that cut score had been set to maximize classification accuracy for a much smaller sample of normally achieving and exceptional pairs who had been matched on race/ethnicity, age, gender and grade. For this much larger, non-matched sample, a cut score of 11.5 was associated with the best combination of sensitivity and specificity. This particular score was selected because if one moved the cut up to include one more exceptional child, which would raise the test’s sensitivity to 86%, specificity would have declined to 76%, a level that is below the minimum level of classification accuracy recommended by Meisels (as cited in Bracken, 1991, p. 496). Shown in Table 1 is an adjusted form of that distribution. Because scores achieved ranged from a -29 to +40, the scores were grouped into sequential sets of approximately 10 points each. As can be seen, the cut score of 11.5 was associated with a sensitivity of 81% (17 of 21) and a specificity of 81% (143 of 177). However, the cost of correctly classifying children with mild learning problems was, according to these data, a high rate of over referral. Sixty-seven percent (34 of 51) of children with scores below the cut were from the normally achieving sample.

TABLE 1

Frequency Distribution of Number of Children in the Normally Achieving and Exceptional Groups As a Function of Their Screening Score

<table>
<thead>
<tr>
<th>Screening Score</th>
<th>Normally Achieving</th>
<th>Exceptional</th>
</tr>
</thead>
<tbody>
<tr>
<td>-29 to -20</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>-19 to -10</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>-9 to 0</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>1 to 11</td>
<td>18</td>
<td>5</td>
</tr>
<tr>
<td>Cut Score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11.5 to 20</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>20.5 to 29.5</td>
<td>60</td>
<td>2</td>
</tr>
<tr>
<td>30 to 40</td>
<td>23</td>
<td>0</td>
</tr>
<tr>
<td>177</td>
<td>21</td>
<td></td>
</tr>
</tbody>
</table>

Cross Validating / 171
The positive predictive accuracy of the test (Feinstein, 1976), which in this case is the percentage of all children called exceptional who actually were exceptional, was only 33%.

Interrater Reliability

Two raters scored ninety-six of the data sheets in order to obtain a measure of scoring consistency. Shown in Table 2 is the percentage agreement using two criteria. Only the percentage agreement for the semantic information/visual (fruit) task is unacceptably low. However, when the measure of agreement is plus or minus one point, the interrater reliability for this task is quite good.

Assessing the Impact of Testing Language, Race/Ethnicity and Gender on Screening Test Performance

Only children in the normally achieving sample were used to evaluate impact of these variables on performance because in practice, the screening test would be administered to preschool children not yet receiving special education services.

Testing language. The first analysis included all children in the normally achieving sample (N = 177). The screening test performance of the children who were tested in English (n = 163) versus those who were tested in either Spanish alone or both English and Spanish (n = 14) was examined. Mean performance of children in these two groups is shown in Table 3. It can be seen that the mean screening score was much higher for those children tested in English when compared to the children who were tested in Spanish or both languages (18.8 vs. 8.5 respectively). This difference favoring the English only group was significant, t(175) = 3.18, p < .01.

A second analysis was computed in order to see if the testing language effect was still present when only Hispanic children (Black or White) were considered, the only groups within which different testing languages were required. As was true for the larger group, the Hispanic children tested in English (n = 80) achieved a significantly higher mean score (17.3) than the 13 Hispanic children who were tested in Spanish or both languages (7.7). Indeed, the mean performance of the latter group was below the cut score of 11.5 suggesting that more of the Hispanic children who could not be tested in English achieved scores that fell below the cut. This implication was examined by computing a frequency distribution of the number of Hispanic children achieving each screening score as a function of testing language. As can be seen in Table 4, more than half (7 of 13) of the Hispanic children who could not be tested in English had scores that would indicate risk for subsequent academic problems, while only 22% (18 of 80) of the Hispanic children tested in English achieved scores that fell below the cut. This difference in proportion of children in the two language testing groups that obtained scores below versus at, or above, the cut was evaluated with a Chi-square analysis. A significant result, χ²(1, N = 93) = 5.6, p < .02, reflected the observed finding that the proportion of children below the cut off score varied based on testing language. Specifically, there were fewer children than expected with scores below the cut for the Hispanic children tested in English and a greater number of children than was expected with scores below the cut for those Hispanic children who were in the Spanish/both group.

TABLE 2

Percentage Agreement Between Scorers on the Five Tasks in the Preschool Screening Test

<table>
<thead>
<tr>
<th>Tasks</th>
<th>Percentage Agreement</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Exact</td>
<td>Within 1 Point</td>
</tr>
<tr>
<td>Taxonomic Generation</td>
<td>93%</td>
<td>99%</td>
</tr>
<tr>
<td>Picture Pointing</td>
<td>90%</td>
<td>98%</td>
</tr>
<tr>
<td>Semantic Information/Verbal</td>
<td>80%</td>
<td>99%</td>
</tr>
<tr>
<td>Picture Rhyme</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Semantic Information/Visual</td>
<td>68%</td>
<td>94%</td>
</tr>
</tbody>
</table>

Age as a potentially confounding variable. One would expect children’s cognitive performance to correlate positively with chronological age. Therefore, a comparison of the two testing language groups in terms of their chronological age was conducted to determine if these two groups, formed on a post hoc basis
to examine the impact of testing language, differed on this potentially confounding variable. A simple t-test was computed, contrasting the two language testing groups in terms of their chronological age at the time of testing, first using the full sample of all normally achieving children (see Table 3). The mean chronological ages of the English only sample (56.7; SD = 4.2) and that of the Spanish/both sample (54.6; SD = 4.3) were not significantly different, t(175) = 1.80, p = .05.

A similar difference in chronological age was present for the Hispanic only testing language contrast. The mean age of the 80 Hispanic children tested in English (56.5; SD = 4.1) was less than two months greater than the mean age of the Hispanic sample who were tested in Spanish or both languages (54.9; SD = 4.2). This difference was not significant either, t(91) = 1.30, p = .10.

**Translation as a potentially confounding factor.**

As a final check on potential confounding variables of the testing language effect, instructions in Spanish were back-translated into English. The Spanish, back-translated into English instructions were consistent with those administered in English. Poorer performance of Hispanic children who had to be tested in Spanish or both English and Spanish was not due to the presence of imperfect or inaccurate instructions.

**Race/Ethnicity.**

The initial contrast of racial/ethnic groups utilized all the groups except for children designated by their parent(s)/guardian as Black/Hispanic (n = 6) or Asian (n = 2) because there were too few children in each. Thus, the impact of race/ethnicity was first examined across five groups; White/non Hispanic (n = 36), White/Hispanic (n = 87), African American (n = 16), Haitian (n = 18) and multi-racial (n = 12). Although the mean performance of the groups varied from 22.4 to 16.0, the overall ANOVA was not significant, F(4, 164) = 2.01, p > .05.

Since in previous studies (e.g., Scott, Fletcher, et al., 1998; Scott, Deuel et al., 1998) racial/ethnic performance differences have typically contrasted fewer groups, for example White/nonHispanic versus Black versus White/Hispanic, the children in this study were first reclassified into a White/nonHis-

---

**TABLE 3**

Mean Screening Score of All the Normally Achieving Children and the Hispanic Only Normally Achieving Groups as a Function of Testing Language

<table>
<thead>
<tr>
<th>All Normally Achieving</th>
<th>Hispanic Normally Achieving</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>English</strong> (n = 163)</td>
<td><strong>Spanish/Both</strong> (n = 14)</td>
</tr>
<tr>
<td>18.8 (SD = 11.5)</td>
<td>8.5 (SD = 12.6)</td>
</tr>
<tr>
<td><strong>English</strong> (n = 80)</td>
<td><strong>Spanish/Both</strong> (n = 13)</td>
</tr>
<tr>
<td>17.3 (SD = 12.7)</td>
<td>7.7 (SD = 12.7)</td>
</tr>
</tbody>
</table>

---

**TABLE 4**

Frequency Distributions of Number of Normally Achieving Hispanic Children Tested in English and Number of Normally Achieving Hispanic Children Tested in Spanish or Both English and Spanish Who Achieved Each Score

<table>
<thead>
<tr>
<th>Screening Score</th>
<th>Testing Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>−20 to −16</td>
<td>English</td>
</tr>
<tr>
<td>−15.5 to −11</td>
<td>Spanish/Both</td>
</tr>
<tr>
<td>−10.5 to −06</td>
<td>1</td>
</tr>
<tr>
<td>−5.5 to −01</td>
<td>2</td>
</tr>
<tr>
<td>−0.5 to 05</td>
<td>2</td>
</tr>
<tr>
<td>5.5 to 11</td>
<td>7</td>
</tr>
<tr>
<td>Cut Score</td>
<td>80</td>
</tr>
</tbody>
</table>

| 11.5 to 15      | 12               |
| 15.5 to 20      | 12               |
| 20.5 to 25      | 17               |
| 25.5 to 30      | 13               |
| 30.5 to 35      | 6                |
| 35.5 to 40      | 2                |
| 80               | 13               |
panic group \((n = 36)\) and a single minority group consisting of all the other normally achieving children who by national demographic data are considered to belong to minority groups \((n = 141)\) and a one way ANOVA of racial/ethnic groups was computed. Mean screening score of the White/nonHispanic group \((22.4, SD = 7.6)\) was significantly higher, \(F(1, 175) = 6.45, p < .02\), than the mean performance of the omnibus minority group \((16.9, SD = 12.5)\).

The 13 children in the Hispanic group who had been tested in Spanish or both languages had a much lower score \((7.7)\) than the Hispanic children tested in English \((17.3; see Table 3)\). Performance of the Hispanic group \((16.9, SD = 12.1)\), \(F(1, 161) = 4.7, p < .04\), that consisted only of children tested in Spanish. There was no obvious solution to “Where should she be placed?” for this analysis, hence the exclusion. Results were the same as were previously reported. The 35 children in the White/nonHispanic group evidenced a higher mean level of performance \((22.5, SD = 7.7)\) than the 128 children in the omnibus minority group \((17.8, SD = 12.1)\), \(F(1, 161) = 4.7, p < .04\), that consisted only of children tested in English.

One final analysis of racial/ethnic effects was conducted. This compared the performance of the White/nonHispanic group \((n = 5)\) with that of the Hispanic children (White or Black) tested in English \((n = 80)\) and Hispanic children (White or Black) tested in Spanish or both languages \((n = 13)\). As was found before, the White/nonHispanic group had the highest mean score \((22.5; SD = 7.7)\) followed by the Hispanic children who were tested in English \((17.3; SD = 12.7)\) while the Hispanic children tested in Spanish or both languages had the lowest mean performance \((7.7; SD = 12.7)\). The ANOVA contrasting these three groups was significant, \(F(2, 125) = 7.9, p < .01\). Follow-up \(t\) tests indicated that all three groups differed significantly from one another, all \(p's < .03\).

**Gender.** Although the mean score for females was higher \((19.8, SD = 11.0)\) than the mean score for males \((16.4, SD = 12.4)\), a \(t\) test comparing these mean differences was not significant, \(t(175) = -1.9, p > .05\).

**Discussion**

This was a cross-sample validation of a new preschool screening test whose content had been selected based on the characteristics of a smaller, matched sample (Scott, Fletcher, et al., 1998). The screening test did demonstrate cross-sample validity when administered to a much larger, nonmatched sample. As expected, there was some reduction in the absolute level of classification accuracy, but both sensitivity and specificity were still above 80%, the minimal level of sensitivity and specificity for a screening test suggested by Meisels (as cited in Bracken, 1991, pg. 496). This level of concurrent predictive accuracy is particularly impressive when one considers the fact that none of the exceptional children in this sample were classified as EMH. Thus, the expected intellectual and hence cognitive difficulties of the exceptional group were much less than they were in previous samples that did include children classified as EMH.

Accurate identification of this mildly impaired group came at the cost of a high rate of over referral (67%). However, if one considers these results in terms of the referral rate associated with the normally achieving sample, the type of children the screening test would be administered to, a different perspective emerges. Lichtenstein and Ireton (1984, p. 206) have suggested that as “...a general rule of thumb,” a referral rate of 1 1/3 to 2 1/2 times the base rate is needed in order to identify children in low base rate groups, that is base rates of from 3% to 10%. The best estimate of the base rate for our mildly impaired target group based on the most recent data (2000-2001 school data) was 6% which was the sum of the base rates for children classified as either EMH or LD. Since this base rate is at the lower end, we used the higher end of the range for referral rates. Two and one half times the base rate of 6 yields a
needed referral rate of 15%. This is not too far from the referral rate for the normally achieving sample observed in this study, which was 19%. In evaluating this referral rate, one must further consider that some of the normally achieving preschool children with scores below the cut will demonstrate learning problems in the next few years. For them, this initial identification will prove to be correct and the referral appropriate; they were at risk for subsequent educational problems. A long-term predictive study is necessary to examine that issue. Such studies have been advocated as part of the evaluative process for screening tests as they, “... provide much needed research exploring the link between screening decisions and future diagnostic/placement decisions” (Lenkarski, Singer, Peters, & McIntosh, 2001, p. 23).

Some have argued that the positive predictive accuracy associated with a screening test is the index that school personnel consider to be the most important (Gredler, 1997). Others, while agreeing that this particular psychometric characteristic is important, have taken the position that a test’s sensitivity is the most important characteristic (e.g., Limbos & Geva, 2001; Rafoth, 1997) since the purpose of a screening test is to identify all, or nearly all, children who are at risk. They would argue that subsequent to the referral process, more extensive educational evaluations will confirm or rule out the presence of a disability, but that identifying all, or most, children who are at risk for academic problems is the most critical goal. By that argument, the cross-sample validation was highly successful since 81% of the exceptional sample was correctly identified.

The cut score used in this study, 11.5, was much lower than the cut score of 23 that maximized sensitivity and specificity with the smaller, matched sample. While we would stipulate that this cut score of 11.5 is the one to be used with this screening test, since it was established using a large sample of children, none of whom were matched to any exceptional child, one would still want to conduct another cross-sample validation to check levels of classification accuracy achieved with a new sample using this fixed cut score.

Importance of the particular cut score associated with a screening test may, however, not be that critical. Lenkarski et al. (2001, p. 23) have said, “... it would be essential for clinicians to determine the appropriate cut-off score for the specific population of children they are screening. Ideally, the determination of an appropriate cut-off score should be based upon locally developed norms.” If this methodology were applied to this screening test, then the fact that it was possible to set a cut score with this particular set of tasks that resulted in high levels of sensitivity might be the more important component of this cross-sample validation. One would, however, expect that this cut score should hold to identify preschool children with mild learning problems who come from a population where national minorities make up the majority, a sample that would be comparable to the one in this study. Were another cross-sample validation done in this same county, the cut of 11.5 would be the one applied and the levels of sensitivity and specificity obtained would be the values expected in its use with similar populations.

Interrater consistency was good for all but the semantic information task. However, the finding that marked improvement in interrater consistency was observed when a score within 1 point was considered (see Table 2), suggests that changes in the instructions described in the scoring manual may reduce these scoring disagreements dramatically. Actually, modifications in scoring instructions made after the Scott, Fletcher, et al. (1998) study, resulted in large improvements in the percentage of exact agreements on two of the generating tasks. Interrater agreement increased from 78% to 93% for the taxonomic generation task and from 64% to 80% for the semantic information/verbal task. An evaluation of scoring errors on the semantic information/visual task should lead to similar levels of improvement in scoring of that task.

The depressed screening test performance associated with preschool children who needed to be tested in Spanish, or both English and Spanish, is a finding that was seen previously with another large sample of normally achieving children. Scott, Deuel, Sanchez, and Levine (1995) reported a significantly higher screening test score for those Hispanic children tested in English compared to those Hispanic children tested in Spanish/
both. Perhaps the most dramatic impact of testing language reported in Scott et al., however, was seen when one examined only those normally achieving children whose screening scores were in the (approximately) lowest 8% of the sample and were, therefore, predicted to require special education services by third grade. Of the Hispanic children predicted to require special education services, nearly all (20 of 21) had been tested in Spanish or both English and Spanish. This degree of impact was not seen in the present study but proportionally, there were significantly more children from the Spanish/both language testing group (54%) than from the English testing group (22%) with scores that fell below the cut.

How valid are the results of a screening test for children who cannot be tested in English? Do the poorer screening scores, over half falling below the cut, reflect a mild cognitive impairment that is also manifest in their inability to acquire a competency in English equal to their same ethnic and age peers, or is it simply a reflection of, for example, the differential exposure of these particular children to English, that might only have a temporary effect on their performance? A long-term predictive study would provide answers to that question by, for example, examining achievement scores of these two groups of children in grade school. If significant differences were present, that would imply that the testing language effect was reflecting true differences in cognitive capabilities rather than a temporary difference in acculturation or English language exposure. Until one knows, the conservative decision regarding preschool children who must be tested in their native language might be to closely monitor any child belonging to this group who achieves a score below the cut, and to retest when that child can be tested in English or if that child shows any other signs of learning problems.

The race/ethnicity effect, reflecting the higher performance of the White/nonHispanic group compared to the omnibus minority group (compared to all or just those tested in English) was also reported in the Scott et al. (1995) study where the White/nonHispanic group had significantly higher scores than either the Hispanic or African American groups, who did not differ. Additionally, significant differences favoring White/nonHispanic children on the screening battery over minority groups was observed in Scott, Fletcher, et al. (1998). This pattern of differences was present for both the preschool and kindergarten battery, although only significant at the preschool level. Given known demographics, however, one would like to be able to examine these racial/ethnic differences in light of a possible confound of race/ethnicity with socio-economic status. Such an examination is critical to an understanding of what underlies these putative racial/ethnic differences in screening performance.

Although this difference is persistent, it does not appear to be as great as that related to testing language where the Hispanic children who were tested in Spanish or both languages had a mean screening score that was below the cut. Actually, no racial/ethnic group had a mean screening score that fell below the cut except for those Hispanic children who were tested in Spanish/both. Indeed, when the Hispanic sample was separated into those tested in English and those tested in Spanish/both, the former showed a significantly higher screening performance ($M = 17.3$) than the latter ($M = 7.7$), although it was still significantly below that of the White/nonHispanic group ($M = 22.5$). Clearly, testing language is a critical variable to consider when screening young Hispanic children for the presence of mild learning problems or general cognitive competency.

This screening test is still in the developmental stage. However, one is encouraged by the cross-validated levels of sensitivity and specificity that were observed (albeit with a change in cut score), the marked improvements in interrater reliability, the absence of significant gender effects and the demonstrated test-retest reliability. Clearly further work is required. Cumulative results, however, suggest that it may well be possible to identify most young children with mild learning problems and consequently to provide early educational interventions for them before they experience years of school failure. Early identification and intervention should minimize the impact of their mild cognitive problems on their subsequent academic performance.
References


Received: 24 April 2002
Initial Acceptance: 10 June 2002
Final Acceptance: 15 October 2002