Development of the Miller Screening for Preschoolers

Lucy Jane Miller

Key Words: child development disorders • tests, by title, Miller Screening for Preschoolers

This paper describes the development and initial validation of the Miller Screening for Preschoolers (MSP). A pilot edition of the test was administered to 174 preschool children in the state of Colorado. Of these, 39 had been previously identified as developmentally delayed. On the basis of preliminary analyses, test items were discarded if they were insensitive to differences between the at-risk group and the not-at-risk group or if they were nonsignificantly related to developmental age trends; if they were only weakly related to the domain total score; if they were redundant with other items within a domain; if they were too easy or too difficult for any given age; if they were rated low by the testers on administration and scoring issues; or if the test materials were deemed to be too expensive. This paper presents final statistics for the selected screening test items and discusses implications for the design, construction, and validation of developmental tests.

Lucy Jane Miller, PhD, OTR, is President, Developmental Technologies, 8101 East Prentice Avenue, Suite 518, Englewood, Colorado 80111.

This article was accepted for publication December 12, 1988.
screening that resulted in the development of the Miller Assessment for Preschoolers (MAP) (Miller, 1987, 1988a, 1988b, 1988c; Miller & Lemerand, 1986; Miller & Schouten, 1988; Widerstrom, Miller, & Marzano, 1986). The MSP was designed to be used as an independent source of screening information or as a companion to the MAP (Miller, 1988c). Although the MSP and the MAP were intended to assess the same skill domains, there is no item overlap between the two tests.

The MAP is more comprehensive than the MSP. It could potentially be used to substantiate indications of risk indicated by the shorter MSP. The MAP requires relatively expensive test materials, a high level of clinical expertise, and at least a half hour to administer. The MSP was developed as a shorter, more economical alternative to the MAP. A preliminary analysis of a large pool of items identified the best of the MSP items.

**Method**

**Subjects.** The subjects were 174 preschool children (91 boys and 83 girls). Sites representative of the population of Colorado or of major demographic characteristics were selected. Fifteen field testers randomly selected locations from all possible sources of children in their locale. The sample was selected with a random numbers table, with no more than 5% of any one site being tested. The subjects' ages ranged from 35 months to 73 months (mean = 52.84 months, SD = 10.75). Four major ethnic groups were represented (Caucasian, 74%; Black, 13%; Mexican-American, 10%; Oriental, 3%). Although drawn entirely from the state of Colorado, the sample was heterogeneous in terms of family demographics such as educational level, vocational status, and income. Of the 174 children, 39 were identified as developmentally delayed on the basis of Colorado State rules and regulations describing an at-risk child as one who has a 25% delay in two or more functional areas. The pattern of developmental deficits varied between children and included language, motor, cognitive, and behavior problems.

**Instrument.** The initial pool of 176 items considered for inclusion on the MSP came from items field-tested but not included in the final edition of the MAP (Miller, 1985, 1988c). Because of space and time limitations, none of these items were included on the MAP, although many had excellent item characteristics. All 176 items demonstrated excellent psychometric properties as well as clinical validity and utility during the tryout. The group of 176 items could be combined into 22 composite scores.

**Examiners and data collection.** The children were tested individually by an occupational therapist, a speech therapist, or a school psychologist. All of the examiners held at least a master's degree and had previous experience in developmental assessment. A 3-day training session was provided to ensure uniform and reliable test administration and scoring procedures. Procedural reliability checks were performed until each examiner achieved at least a 95% accuracy rate in administration. In addition to administering and scoring the tests, the examiners completed a 20-page questionnaire that used open-ended questions and ratings scales related to task characteristics (e.g., ease and reliability of administration and scoring, clinical validity, and appropriateness for children) to elicit their feedback on the testing procedures. A content analysis of the examiners' responses was performed, and averages for ratings on specific administration and scoring issues and endorsements were calculated. Although not reported here, these data were as important as psychometric considerations in the final item-selection process.

For the sake of brevity, this study focuses on those items that were ultimately selected in accordance with the following criteria: (a) ability to differentiate between the at-risk group and the not-at-risk group, (b) sensitivity to developmental age trends, in which increased age is associated with higher pass rates on the items, (c) substantial contributions to their respective domains, (d) nonredundancy with other items within a domain, (e) pass-fail rates sufficient to demonstrate individual differences in ability, and (f) favorable ratings from interviewing clinicians on administration and scoring issues and on time- and cost-effectiveness.

A preliminary analysis eliminated more than half of the items from the original pool because of psychometric or practical criteria. The results that follow refer to the final statistics for those items that were retained.

**Data analysis.** Composite scores were created by linear combinations of similar types of items to simplify the data. For example, the Object Memory task consisted of four pass-fail trials. Scores on the individual trials were combined to provide a single Object Memory task composite score. If an item was scored with more than one metric, the scores were standardized by z score transformations and combined into a single score. For example, the times required to complete the measures for six successive Problem Solving tasks were combined and scored as a single mean item score. The composite scores were grouped into one of four developmental domains: language, motor, cognitive, or behavior.

---

1 Complete correlation matrixes and summary tables for the preliminary analysis are available from the author upon request.
Results

Reliability. Coefficient alpha (Cronbach, 1951) was calculated for the four domains for three age groups (32-44 months, 45-56 months, and 57-73 months) and for the total sample (see Table 1). The coefficients indicated acceptable levels of reliability for each domain and stability across age groups. All internal consistency coefficients for the full-scale MAP Screen for this sample approached or exceeded .90.

Test-retest data were collected for 38 randomly chosen children who were retested on all of the items after a 1-2 week interval. The values for the language, motor, cognitive, and behavior domains were .87, .98, .71, and .95, respectively. The relatively low test-retest reliability for the motor items warrants further attention. Motor functions appear to be either generally more variable than functions in other areas or perhaps more susceptible to practice effects.

Twelve children were randomly selected for the interrater reliability study. One examiner independently scored the performance of a child who was tested and scored by one of four other examiners. Pearson product-moment correlations were calculated to determine the relationship between the two sets of test scores. The values were .99 for both the language and the motor items, .93 for the cognitive items, and .79 for the behavior items. These data indicate generally high levels of interrater reliability.

Group differences. Separate univariate analyses of variance (ANOVA) were performed to compare the children who were at risk and the children who were not at risk. Separate tests were performed for the test tasks to determine the comparative discrimination power of each. Effect sizes were calculated for each on the basis of univariate F-ratios (Cohen, 1969). The results of these analyses are shown in Table 2. The two groups were significantly differentiated on all but three areas: Association, Object Memory, and Problem Solving, including Number Correct and Time to Complete. Although the results for these measures fell short of statistical significance, the group means were in the expected direction. In addition, effect sizes were generally large (range = .36 to 2.22; median = 1.02), that is, exceeding one standard deviation (Cohen, 1969). These data indicated high power of discrimination for the new screening test with respect to the group criterion.

Developmental-age trends. The sample was divided into six age groups: 33-38 months, 39-44 months, 45-50 months, 51-56 months, 57-62 months, and 63-73 months. Separate univariate ANOVAs were performed to compare age groups. As in the analysis of group differences, separate tests were conducted and effect sizes were calculated for each test task to determine the comparative discrimination power of each. The results of these analyses are shown in Table 3. Strong linear age trends were evident on all measures, and all trends were in the expected direction, such that increased age was associated with higher pass rates or increased efficiency (i.e., less time to complete tasks). Effect sizes were calculated to evaluate the strength of the age-performance relationship for individual items. The values ranged between .33 and .84 (median = .55), indicating moderate to large effects (Cohen, 1969) and demonstrating that the screening test is sensitive to developmental-age trends.

Ethnic differences. An ANOVA comparing the ethnic groups' mean item scores was performed to determine the differential bias of the test items. Overall, there was little evidence of performance differences between the ethnic groups. However, a significant effect due to ethnicity was obtained for the Problem Solving Number-Correct measure [F(3, 126) = 4.12, p < .008] and for the Problem Solving Time to Complete measure [F(3, 128) = 4.39, p < .006]. A post hoc comparison with Duncan's Multiple Range test showed a similar pattern of differences on both measures, with the Hispanic children, on the average, performing better than the Black children (for both, p < .05). This is an unexpected finding that calls for further attention.

Discussion

Developmental and preschool assessment has been marked by a number of technical problems (Bracken, 1987; Lidz, 1986). These problems reflect the lack of a comprehensive model of test development and the difficulties inherent in obtaining accurate measures of young children's emerging skills. This state of affairs is a source of concern in view of the passage of Public Law 99-457, which will result in a substantial increase in the number of children who undergo preschool screening. The value of screening procedures to meet current demand is directly related to their usefulness in detecting handicapping disorders, in evaluating school readiness, and in predicting future learning problems.
The conclusion that the MSP could serve as a useful screening tool was supported by the evidence collected in this study. Of special note were the substantial differences between the children with developmental delays and the children without developmental delays as well as the consistent developmental-age trends. The present results indicate that the MSP is sensitive to individual differences in developmental status, some of which would be difficult to ascertain except through in-depth testing, which would require a high level of clinical expertise. Further, the assessment exhibited little cultural bias, with negligible ethnic differences. Taken together, these data indicate that the pilot edition of the MSP combines accuracy, efficiency, content relevance, and freedom from item bias.

However, the present findings must be substantiated. The significance of these results is qualified by the large number of statistical tests performed on a relatively small sample and the specific risk criteria used to select subjects for the at-risk and not-at-risk groups. The findings must be replicated and extended with larger samples and with more clearly defined risk groups to determine their stability and generality. In addition, follow-up studies will be required to evaluate the predictive efficiency of the MAP Screen with respect to later functioning. The instrument must be validated against external criteria such as later physical development, psychological adjustment, and school status. Predictive validity data are also needed to provide a basis for final scaling and scoring decisions. It will be especially important to determine the optimum cutoff scores that identify children who may be at risk for future problems. This information will ultimately aid practitioners who use the test.

### Conclusion

A screening outcome is not equivalent to a diagnosis of neurological abnormality, learning disability, or psychopathology. However, as a possible basis for referring an at-risk child for more comprehensive testing, which could in turn lead to early intervention, screening results can represent an important first step in the early identification and remediation of childhood developmental disabilities, because screening...
Table 3
Analysis of Variance Comparing Six Age Groups: Pilot Edition of the Miller Screening for Preschoolers

<table>
<thead>
<tr>
<th>Age Group</th>
<th>33-38 (n = 17)</th>
<th>39-44 (n = 33)</th>
<th>45-50 (n = 23)</th>
<th>51-56 (n = 23)</th>
<th>57-62 (n = 43)</th>
<th>63-73 (n = 34)</th>
<th>F</th>
<th>p</th>
<th>Effect Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LANGUAGE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sentence Repetition</td>
<td>0.20</td>
<td>0.22</td>
<td>0.39</td>
<td>0.70</td>
<td>0.77</td>
<td>0.85</td>
<td>10.89</td>
<td>.00001</td>
<td>0.60</td>
</tr>
<tr>
<td>Fluency</td>
<td>2.14</td>
<td>2.89</td>
<td>6.41</td>
<td>7.91</td>
<td>8.54</td>
<td>12.79</td>
<td>23.22</td>
<td>.00001</td>
<td>0.74</td>
</tr>
<tr>
<td>Association</td>
<td>0.67</td>
<td>0.69</td>
<td>0.77</td>
<td>0.69</td>
<td>0.75</td>
<td>0.86</td>
<td>3.95</td>
<td>.003</td>
<td>0.39</td>
</tr>
<tr>
<td>MOTOR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Balance</td>
<td>4.05</td>
<td>5.26</td>
<td>7.28</td>
<td>6.20</td>
<td>8.61</td>
<td>11.03</td>
<td>13.12</td>
<td>.00001</td>
<td>0.57</td>
</tr>
<tr>
<td>Pattern Alternation</td>
<td>0.35</td>
<td>0.48</td>
<td>0.64</td>
<td>0.81</td>
<td>0.80</td>
<td>0.87</td>
<td>7.37</td>
<td>.0001</td>
<td>0.49</td>
</tr>
<tr>
<td>Number Correct</td>
<td>22.36</td>
<td>20.45</td>
<td>15.34</td>
<td>11.38</td>
<td>11.69</td>
<td>10.04</td>
<td>13.28</td>
<td>.00001</td>
<td>0.63</td>
</tr>
<tr>
<td>Time to Complete</td>
<td>0.69</td>
<td>1.65</td>
<td>4.45</td>
<td>4.28</td>
<td>5.04</td>
<td>7.89</td>
<td>25.54</td>
<td>.0001</td>
<td>0.79</td>
</tr>
<tr>
<td>Motor Accuracy</td>
<td>17.15</td>
<td>11.43</td>
<td>10.63</td>
<td>9.43</td>
<td>6.96</td>
<td>4.66</td>
<td>4.53</td>
<td>.0006</td>
<td>0.33</td>
</tr>
<tr>
<td>Motor Planning</td>
<td>10.24</td>
<td>13.07</td>
<td>21.65</td>
<td>21.83</td>
<td>26.79</td>
<td>34.29</td>
<td>29.52</td>
<td>.0001</td>
<td>0.84</td>
</tr>
<tr>
<td>Time to Complete</td>
<td>1.37</td>
<td>2.04</td>
<td>2.68</td>
<td>2.88</td>
<td>3.50</td>
<td>3.88</td>
<td>18.68</td>
<td>.0001</td>
<td>0.67</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.66</td>
<td>0.64</td>
<td>0.72</td>
<td>0.69</td>
<td>0.73</td>
<td>0.78</td>
<td>6.63</td>
<td>.0001</td>
<td>0.45</td>
</tr>
<tr>
<td>Stereognosis</td>
<td>0.66</td>
<td>0.74</td>
<td>0.82</td>
<td>0.91</td>
<td>0.92</td>
<td>0.99</td>
<td>7.72</td>
<td>.0001</td>
<td>0.44</td>
</tr>
<tr>
<td>COGNITIVE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Closure</td>
<td>0.64</td>
<td>0.74</td>
<td>0.82</td>
<td>0.91</td>
<td>0.92</td>
<td>0.99</td>
<td>7.72</td>
<td>.0001</td>
<td>0.44</td>
</tr>
<tr>
<td>Object Memory</td>
<td>0.67</td>
<td>0.71</td>
<td>0.89</td>
<td>0.91</td>
<td>0.88</td>
<td>0.96</td>
<td>4.49</td>
<td>.0009</td>
<td>0.38</td>
</tr>
<tr>
<td>Quantitative</td>
<td>0.54</td>
<td>0.63</td>
<td>0.67</td>
<td>0.77</td>
<td>0.78</td>
<td>0.90</td>
<td>8.98</td>
<td>.0001</td>
<td>0.49</td>
</tr>
<tr>
<td>Problem Solving</td>
<td>0.24</td>
<td>0.25</td>
<td>0.52</td>
<td>0.54</td>
<td>0.67</td>
<td>0.71</td>
<td>11.78</td>
<td>.0001</td>
<td>0.53</td>
</tr>
<tr>
<td>Number Correct</td>
<td>42.76</td>
<td>41.65</td>
<td>32.94</td>
<td>31.35</td>
<td>29.45</td>
<td>25.73</td>
<td>15.27</td>
<td>.0001</td>
<td>0.60</td>
</tr>
<tr>
<td>Time to Complete</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Results can have far-reaching effects, the screening process must have technically sound and appropriate measurement tools.

The MSP has satisfied important test-construction criteria. Specifically, the results of this pilot study provide preliminary evidence of the test's potential as a valid and reliable brief and cost-effective developmental screening test. The use of this instrument warrants further investigation.

Acknowledgment

Research for this study was supported by Grant 1R43 HD 2267-01 from the U.S. Department of Health and Human Services (Human Learning and Behavior), Small Business Innovation Research Program, Washington, DC.

References


Education of the Handicapped Act Amendments of 1986 (Public Law 99-457), 20 U.S.C. §1400. The MSP has satisfied important test-construction criteria. Specifically, the results of this pilot study provide preliminary evidence of the test's potential as a valid and reliable brief and cost-effective developmental screening test. The use of this instrument warrants further investigation.

Liden, C. S. (1986). Preschool assessment: Where we have been and where we are going. Special Services in the Schools, 2, 141-158.


**Wanted: Candidates for AJOT Editorial Board**

We need people to serve as reviewers for manuscripts submitted to *AJOT*. Desirable traits are professional expertise, familiarity with occupational therapy literature, knowledge of research design and statistics, knowledge of occupational therapy interventions, and willingness to respond to requests for reviews in a timely manner. We are also seeking individuals who can serve as statistical consultants on manuscripts.

Please send qualifications, areas of interest, and a brief résumé to Elaine Viseltbear, Editor, 616 Tanner Marsh Road, Guilford, Connecticut 06437.