A Critique of the Standardization of the Miller Assessment for Preschoolers

The Miller Assessment for Preschoolers (MAP) is a standardized screening instrument for preschoolers who are at risk for moderate to severe academic problems. A review of the procedures used to standardize the MAP found that it is a well-developed screening instrument that exhibits the qualities of good item discrimination, good test structure and content, good data collection on normal subjects, and good inter-rater and test-retest reliability. The MAP is effective in screening normal children who fall in the lowest 20th percentile. Since the intent of the MAP is to screen preschool children with moderate to severe pre-academic problems, further reliability and validity studies described in this paper need to be conducted before the MAP can be used effectively in screening children with delays. Caution should be used in categorizing normal children as "high risk" or suspect until decision validity studies have been completed.

The Miller Assessment for Preschoolers (MAP) (1) was developed by Lucy Miller, an occupational therapist, in response to a need for a standardized screening instrument for preschoolers who are at risk for moderate to severe academic problems. The test is designed for use with children ranging in age from 2 years, 9 months to 5 years, 8 months and can be administered by a variety of clinical and educational personnel. The test measures skills within the domains of sensory and motor abilities, speech and language, cognition, behaviors, and visual-motor integration. The intent of this paper is to critique the standardization procedures used in the MAP. Guidelines set forth by Berk and DeGangi (2) were used in evaluating the reliability and validity of the MAP.

Subjects

The MAP was administered to 1,014 normal children ranging in age from 2 years, 9 months to 5 years, 8 months. A random and stratified sampling procedure was used in selecting children for the test. Fairly equal distributions were obtained for the six different age groups: for sex, region of the United States, and for socioeconomic status, which were relatively close to U.S. Census Bureau figures. Although age differences were considered in the data analyses, race and sex differences were not examined. It may not have been possible to examine race and sex differences since the number of children within each subgroup may have been too small to make judgments. Because evidence shows that Black children from lower socioeconomic families are more advanced in their gross motor skills, whereas Caucasian children tend to excel in language and perceptual tasks in the preschool years (3), it would have been helpful to therapists when screening minority children to know whether such differences were also found on the MAP.

Georgia A. DeGangi, M.S., OTR, is a Clinical Instructor in the Department of Pediatrics, Georgetown University. She is currently in private practice.
In addition to the data collected on the sample of normal children, the MAP was also administered to 90 children with functional delays to examine construct validity. Children within this sample were defined as those with perceptual, language, or behavioral delays. Children with cerebral palsy, autism, and diagnoses such as mental retardation were excluded. Specific selection criteria should have been defined so the configuration of the sample could be clearly identified. The distribution of age, sex, and race, in addition to the type and extent of delay, should also be presented in the test manual. Without this information it is difficult to determine the effectiveness of the MAP in screening children with pre-academic problems.

Validity

Construct Validity

Decision Validity: The intended purpose of the MAP is to identify preschool children who display moderate to severe pre-academic problems. Since this is the intent, it is necessary to compare a well-defined group of children with pre-academic problems to their normal age mates. Since the sample of delayed children was not well defined and exhibited a wide range of cognitive, language, motor, perceptual, and behavioral deficits, it cannot be determined whether the MAP is effective in screening only specific types of delays (i.e., sensory and motor) or a wider range of delays.

Another major problem that affects the decision validity of the MAP is the arbitrary cutoff points assigned to children who fall into the “at risk” category and those who fall into the “suspect” group. Miller defines those children who are “at risk” as normal subjects who fall below the 5th percentile. Children defined as “suspect” were normal children who fell below the 25th to 34th percentile. Miller recommends that children who are “at risk” should receive intervention and follow-up testing, whereas those who are “suspect” should be re-tested and observed for possible problems. One percent of the normal sample fell below the 5th percentile, whereas 23 percent fell below the 25th percentile. Twenty-three percent is a very high number of normal children who supposedly are being classified as suspect for academic problems. Not only is the method of selecting an arbitrary cutoff point very subjective, but it does not take into account how children with pre-academic problems perform in comparison to a normal sample. Since the intent of the MAP is to screen children with pre-academic problems, then optimal cutting scores should have been determined by comparing the scores of normal children to those of children who display specific deficits so that the best point of group discrimination could be determined (4). This cutting score would then reflect the probability of correctly or incorrectly classifying children either in the false normal or false delayed ranges. In examining the information presented on construct validity on the sample of delayed children, 46 percent of this group fell into the “at risk” group, 29 percent into the “suspect” group, and 25 percent into the normal range. The number of delayed children who actually test within the “at risk” or “suspect” ranges is even lower when age differences are considered. Only 17 to 27 percent of the 2 year, 9 month to 3 year, 8 month level group fell into the “at risk” range and only 30 to 61 percent of the 3 year, 9 month to 5 year, 8 month level groups fell into the “at risk” range. At least 25 to 47 percent of the delayed sample would be falsely classified as normal, and possibly as many as 25 percent of the normal children may be falsely classified as delayed. The false normal rate would be considered to be the most serious type of error in screening children. Since the false normal rate is estimated to be quite high, the usefulness of the MAP as a screening tool for delayed children is highly questionable.

Item Validity: An initial 500 test items were field tested before the 27 items on the current MAP were accepted. The final test items selected were based upon item difficulty (the percent of children passing the item), item discrimination (point biserial correlation of each item), and correlation studies (the relationship among items and the subtest categories). The mean raw score performance for individual test items was computed for the normal subjects. Miller then selected only those items that discriminated among the lowest 20 percent of the normal sample. Examination of the various test items reveals that they also discriminate well across the different age groups. Although only a small number of children were classified as delayed compared to the number in the normal sample size, comparisons were not made between the normal and delayed groups. This comparison is important in order to determine the amount individual test items discriminate between the two groups (5). It appears, then, that the MAP is more effective in screening normal children who fall into the lowest 20th percentile.

Content Validity: How representative the items are in each domain of behaviors is described in a specification table in the Manual. The rationale for these behaviors and for
including the test items designed to measure the specific behaviors is described in detail in Chapter 2 of the Manual. Existing tests were thoroughly researched before test items were created. Testing began in 1973, and more than 800 test items were field tested before the revised test items were finally chosen. Factor analysis was completed for assessing the interrelationships of the test items, which revealed that six primary factors were being tested. A correlation analysis was completed as well that analyzed the items and subtest indices in terms of their contribution to the total test. Test items contributed significantly at the < .01 level, whereas the subtest indices contributed fairly equally (.647 to .778). This information implies that the different subtests measure completely different functions, which they were designed to do. These results provide empirical support for the test structure.

**Domain Validity**

The most important type of validity is the degree of congruence between the test items and the domain of behavior that they are intended to measure. This is achieved by verifying the match between the items and the domain by a panel of judges trained in that domain. If this is not attained, then other test characteristics are meaningless. In the development of test items, Miller consulted with specialists from the fields of occupational and physical therapy, speech and language, psychology, and pediatrics. Unfortunately, a systematic study of domain validity was not conducted in the standardization of the MAP. It is important that it be determined whether other professionals who are expert in the various domains being tested address whether the specific test items measure the behaviors they were intended to measure and whether the collection of items within each domain actually represents that domain.

**Criterion-Related Validity**

In order to test the relationship of results on the MAP to the results of other tests, 30 children were administered the MAP, the Southern California Sensory Integration Tests, Wechsler Preschool and Primary Scale of Intelligence, Illinois Test of Psycholinguistic Abilities, and the Denver Developmental Screening Test. Results of these studies are described in the Manual. Caution is recommended when interpreting the results because of the small sample size. Of particular interest, the MAP was found to identify 24 percent more children as high risk or delayed than the Denver Developmental Screening Test. This may be an indication that the MAP either misclassifies normal children as delayed, or it may reflect the sensitivity of the test in detecting children with delays. This will not be known until further decision and predictive validity studies have been conducted.

**Reliability**

Two types of reliability were examined on the MAP: inter-rater and test-retest. Overall good results were obtained; however, it is thought that further reliability studies should be conducted. Nine registered occupational therapists (OTRs) with pediatric and testing experience were trained in administering and scoring the MAP. These Field Supervisors also partook in the data collection. Since the MAP is designed to be administered by various school personnel, it would be useful to determine whether or not a teacher or an aide who did not have the experience in motor development that an OTR had could also test children reliably using the MAP. It should also be specified in the Manual how much experience in pediatrics and with testing is necessary in order to be a reliable tester.

In the inter-rater reliability studies, 40 normal children were tested. In this type of reliability, one examiner administers the test, while another simultaneously scores, which tells us that a child's performance on a given day can be recorded objectively when one person administers the test and the other observes and scores. It does not tell us that two examiners can administer the same test and yield the same results with a high degree of accuracy. The latter type of reliability is termed interobserver reliability. This is accomplished by two examiners who test the same child 1 to 2 weeks apart. Since the test was designed to screen children with moderate pre-academic problems, it is important to obtain reliability on this group because the observations should be different from those obtained on a normal group. Decision reliability studies should also be conducted to determine whether the number of children tested will be consistently classified as normal or delayed on repeated testings.

Decision reliability differs from test-retest reliability in that samples of both normal and delayed children must be tested on repeated measurements, and classifying the children accurately as normal or delayed consistently is determined. No reliability studies were conducted on the supplemental observations, which included measures of the quality of various responses in language, vision, touch, movement, and drawing. If these observations are to be used for diagnostic
decision making, reliability studies are indicated.

Test-retest reliability studies were conducted on 90 normal subjects who were retested 1- to 4-weeks apart by the same examiner. Three field supervisors were testers. Results were very good, with reliability coefficients ranging from 72 to 94 percent for the subtests, and 81 percent for the total test. The coordination subtest was the only subtest with a reliability coefficient that fell below 80 percent. Three of the items within this subtest (to:er, cage, and vertical writing) had test-retest indices of 69.5 to 76.8 percent. Because of their lower reliability, these three items might possibly have been discarded, or caution should have been expressed in making judgments regarding those particular items on the subtest. Internal reliability was computed at 79 percent for the total test.

Summary
The MAP is a 27-item screening instrument designed for preschool children with moderate-to-severe pre-academic problems. Miller developed the MAP in response to a need for a standardized screening instrument for preschool children. Strengths of the MAP include the following:

1. Data were collected on a large number of normal subjects by using a random and stratified sampling procedure.
2. Fairly equal distributions were obtained within the normal sample for sex, age, region of U.S., and socioeconomic status.
3. The rationale for test item development is well described.
4. Factor analysis and correlational analysis provide empirical support for the test structure and content.
5. The test items discriminate among the lowest 20 percent of the normal sample.
6. Test items clearly differentiate performance across age levels within the normal sample.
7. Good results were obtained in inter-rater and test-retest reliability studies conducted on normal subjects.
8. The MAP is a comprehensive screening tool that measures skills within the domains of sensory and motor abilities, speech and language, cognition, behavior, and visual-motor integration.

Several areas were identified as weaknesses in the standardization of the MAP. Those areas of weakness are discussed below, together with suggestions for future research on the MAP and suggestions for use of the MAP.

1. The sample of delayed children was quite small, compared to the normal sample size. If data could have been collected on 30 to 50 preschool children with delays within each age range, comparisons between the normal and delayed samples would be well based. It is unrealistic to expect a collection of data on as many delayed subjects as there were normal subjects in the MAP standardization sample because of the difficulties in obtaining large numbers of delayed subjects for testing.
2. Selection criteria were poorly defined for the delayed sample and did not include descriptions of the type and extent of delay. If the delayed sample had been well defined, performance characteristics of the delayed subjects could have been described and compared to the performance of normal children. This information could be valuable to a therapist or teacher in educational and therapeutic programming and could also provide a precise description of the sample of delayed children for predictive validity studies, for instance, in determining which types of delays (i.e., motor, language, or overall delay) actually result in academic problems. The task of collecting data for predictive validity studies is certainly not a trivial one and obviously was beyond the scope of the initial standardization project. This author understands that Miller plans to conduct predictive validity studies in the future.
3. Arbitrary cutoff points were used in classifying children as "at risk" or suspect. Only data from the normal sample were used in making decisions about performance. Individual item discrimination and optimal cutting scores should have been determined by comparing the performances of normal and delayed subjects. If a teacher or therapist were to use the data presented in the test manual, there is a 25 to 47 percent chance of misclassifying a delayed child as normal. Since decision validity studies were not conducted, the usefulness of the MAP as a screening tool for children with delays is questionable. It appears that its primary use is for identifying normal children who fall in the lower 20th percentile. It is possible that therapists and teachers who use the MAP would use it only to screen large numbers of normal children in school systems since children with known delays would automatically receive more comprehensive testing, therefore, the fact that the MAP is more effective in screening normal children may not be a major issue at this time. It would be useful to investigate how clinicians are using the MAP to determine the importance of using the MAP for screening delayed children. Again,
predictive validity studies would prove very useful in resolving this question. If children who do indeed exhibit delays are the ones who later display academic problems, and normal children who are identified as “at risk” do not, it becomes critical to improve the standardization of the MAP to screen delayed children as well.

4. Although Miller recommended that normal children falling in the “at risk” range should receive further testing and intervention, it is suggested that more specific criteria should be provided in making a judgment about which children would indeed require intervention. Generally, children who exhibit delays of 1 or more years in gross and fine motor skills, perceptual abilities, or both, require direct therapeutic intervention from occupational therapist, physical therapist, or motor development specialist. Some children, however, may exhibit a wide scatter in performance in the various areas tested and may benefit from consultative services provided by a specialist to the classroom teacher. Construct validity studies should be conducted to determine whether children who fall in the “at risk” range, who do not have delays of over 1 year in various developmental areas, show change in their MAP scores and on other instruments that may demonstrate gains in performance when given direct therapeutic intervention or consultative services provided by a specialist compared to children who do not receive any specific input.

5. Although good results were obtained in the reliability studies, these studies are not considered extensive enough. Since reliability studies were conducted only on normal children, it is difficult to determine whether the MAP could be reliably administered to a delayed child for whom the tool has been designed. Reliability studies were limited only to inter-rater and test-retest and did not include inter-observer or decision reliability studies. These studies could provide information regarding the ability of two observers to obtain the same results when testing on two different occasions, and the ability to correctly classify a child as normal or delayed on repeated testings.

6. No domain validity studies were conducted. This is the most important type of validity, which measures the degree of congruence between the test items and the domain of behavior they are intended to measure. A group of therapists and educational personnel with expertise in the domains measured by the MAP should be systematically interviewed after having an opportunity to use the test to determine whether they feel the specific test items measure what they were intended to measure and whether the collection of items within each domain actually represents that domain.

In summary, the MAP is a well-developed screening instrument that exhibits the qualities of good item discrimination, good test structure and content, good collection of data on normal subjects, and good inter-rater and test-retest reliability. The MAP is effective in screening normal children who fall into the lowest 20th percentile. Test items were specifically designed to discriminate among this group. Categorization of normal children as “high risk” or “suspect” should not be considered valid, however, until decision validity studies have been completed. Since the intent of the MAP is to screen preschool children with moderate to severe pre-academic problems, further reliability and validity studies described in this paper need to be conducted before the MAP can effectively screen children who indeed do have delays. Caution should be used in interpreting results until more data are collected on delayed children. Certainly the investigation of differential performance in terms of sex and ethnicity should also be explored. Given the difficulty of sampling large numbers of delayed children, the task that lies ahead is clearly not trivial. However, the encouraging results that have emerged thus far with the MAP should serve to stimulate new efforts to meet this challenge. Predictive validity studies will contribute immensely toward identifying children with pre-academic problems as early as possible in their development to allow for appropriate therapeutic intervention for those dysfunctions.

REFERENCES