

Test 1

Review of the Process Assessment of the Learner-Second Edition: Diagnostic Assessment for **Math** by LAURA HAMILTON, Senior Behavioral Scientist, RAND Corporation, Pittsburgh, PA:

[Disclaimer: The opinions expressed in this review are those of the review author and should not be attributed to the RAND Corporation.]

DESCRIPTION. The Process Assessment of the Learner-Second Edition: Diagnostic Assessment for **Math** (PAL-II M) is an individually administered assessment of mathematics skills and processes for students in kindergarten through sixth grade (ages 5 through 13 years). It also can be used for older students who are performing at the sixth grade level or lower. It is designed to be administered by skilled personnel such as school psychologists. The user's guide describes several purposes: screening to identify students who may be at risk for low achievement, progress monitoring for struggling students, and diagnosis regarding processing problems. The PAL-II M includes 14 subtests, some of which focus on concepts and skills whereas others measure cognitive or neurodevelopmental processes that research suggests are related to a student's ability to learn **math**. Not all subtests are given at every grade level. Administrators may also choose to use only one or a few subtests that are relevant to specific concerns about a child's **math** learning. Administration of the full test generally requires between 60 and 120 minutes.

For most subtests, scoring can be completed after test administration has ended, though a few require scoring to be done during administration. Item responses are scored either dichotomously or using a partial-credit model in which the maximum score is 2 or 3. Scoring is mostly objective but three subtests involve subjective scoring. A few subtests require the administrator to record the time it took for the student to complete each item, and some have discontinue rules that indicate when to stop testing based on the number of consecutive incorrect item responses. The administration and scoring manual provides clear and detailed instructions for recording and scoring students' responses and for calculating raw and derived scores. The manual also provides guidance regarding what subsets should be administered to address specific concerns about students' **math** learning, as well as recommended interventions for students identified as having various skill deficits.

DEVELOPMENT. According to the test author, the development of the PAL-II M was motivated by a perception that existing intelligence and achievement tests fail to provide information on why a student may fail to demonstrate a skill or what can be done to address students' specific instructional needs. The assessment draws on research on **math** processes and their relationships with achievement, as well as on the test author's teaching and clinical experiences. Its development was guided by a conceptual framework that includes low- and high-level procedural skills and concepts. The PAL-II M was subjected to several stages of testing during its development, including a tryout phase that involved 400 students distributed across Grades K through 6.

TECHNICAL. The standardization was conducted using 225 examiners from across the U.S. who were selected based on experience and access to students in the relevant grade levels. The standardization sample included 700 students, with 100 drawn from each relevant grade level. A stratified random sampling procedure was used, with stratification variables including race/ethnicity, sex, age, region, and parent/guardian education level. Students with certain types of diagnoses, including learning disabilities or language disorders, were excluded. The user's guide presents tables that demonstrate that this sample closely resembles the U.S. Census figures. The administration and scoring manual provides several tables of norms by grade level, including scale score equivalents of raw scores for each subtest.

The test author acknowledges the difficulties inherent in estimating score reliability for tests that have floor or ceiling effects, which is the case for some subtests at some grade levels. In addition, the application of specific approaches to estimating reliability, such as the calculation of internal consistency estimates, is more appropriate for some subsets than for others. The test author has provided extensive information about reliability estimation for each subtest and uses appropriate approaches given the content and scoring systems used in each subtest. For

subtests for which internal consistency was deemed appropriate, split-half and alpha coefficients were calculated. For other measures, test-retest stability was estimated using data from a sample of 129 students who participated in two testing sessions ranging from 2 to 34 days apart. Stability coefficients are correlations between scores obtained from the two sessions and were calculated separately for two groups: Kindergarten through Grade 3 and Grades 4 through 6. The user's guide also includes standard errors of measurement and guidelines for creating confidence intervals, which can be helpful for communicating the degree of uncertainty associated with each score.

Comment [L1]: Reliability: Internal consistency

Internal consistency coefficients of at least .80 were obtained for a majority of subtest and composite scores, though coefficients for a few subtests were slightly lower. Ceiling effects appear to be responsible for low internal consistency estimates for some subtests at some grade levels. Most of the stability coefficients are higher than .70, again with some exceptions. Average scores were higher in the second testing session than in the first session for most subtests, which suggests some practice effects, but the differences tended to be small.

Comment [MG2]: Reliability: Test-retest

Measures of rater agreement were calculated for each subtest and were found to be high for subtests that involved objective rules with minimal interpretation but were lower in cases that required subjective judgment. It is important to point out that the training that the scorers in the standardization process received may have led to higher levels of agreement than would be obtained in operational testing situations.

Comment [L3]: Reliability: internal consistency

Comment [MG4]: Reliability: Interrater

The user's guide presents several sources of validity evidence. It mentions the relevance of consequences to evaluating the validity of test scores, but does not provide any evidence related to consequences of using the PAL-II M.

The discussion of content evidence starts with a discussion of the fact that the content of the PAL-II M is drawn from extensive research linking **math** processes with student achievement, but it does not provide explicit information about the quality and content of the studies on which the assessment is based. Similarly, it states that evidence of children's response processes was gathered to provide additional information about content evidence, but it does not include details about the collection of these data (numbers and ages of students, methods for coding transcripts, etc.).

Comment [MG5]: Validity: Content

Evidence of construct validity consists of intercorrelations among subtest scores for each grade. In general, correlations within domains are higher than correlations across domains. There are several examples of near-zero correlations that appear to be consistent with the expectations of the test author and the nature of the skills measured. For example, Oral Counting and Numeral Writing show no relationships with other subtests in the upper grades, which is not surprising given the fact that most students in the upper grades have obtained a reasonably high level of proficiency on these subtests. Most of the correlations are moderate, suggesting that the various subtests tend to capture different skills, though the correlations are also affected by score reliability, which is lower at the subtest level than for composites or the test as a whole. Concurrent evidence was provided based on correlations between PAL-II M scores and scores on two measures of general cognitive ability: the Differential Ability Scales-Second Edition (DAS-II) and The Wechsler Nonverbal Scale of Ability (WNV). The magnitudes of the correlations varied by subtest and in some cases were fairly low, but because the PAL-II M was designed to measure something other than general ability, these low correlations should not necessarily be viewed as evidence of lack of validity. The test author also presents correlations with the NEPSY-II, a measure of neuropsychological functioning. These correlations varied substantially but tended to be low and in many cases near zero.

Comment [MG6]: Validity: Structural

Comment [MG7]: Validity: relationship with other measures

COMMENTARY. The PAL-II M appears to be a well-researched measure that can provide educators with information that cannot be easily obtained through other sources. The administration and scoring procedures used in the PAL-II M are designed to uncover the processes in which students engage as they attempt each item rather than focusing exclusively on the end result. For example, students use pencils without erasers so that if they make a mistake that they want to correct, they must cross it out rather than erasing it, thereby providing a written record of errors and corrections that can be scored. The user's guide provides clear, detailed information on all aspects of administration and interpretation and offers guidance for working with teachers and others to address students' needs. It also provides various record-keeping forms that can be useful for documenting performance and

communicating with teachers, parents, and others who are involved in a child's educational planning, as well as interview guides, questionnaires, and structured observation forms that can be used to collect supplemental information about student performance.

Although extensive information about score reliability and validity is presented, some additional analyses would be helpful. For example, the standardization sample excluded students with certain special needs but schools may benefit from evidence regarding the appropriateness of the test for these students. It also would be useful to have more information on how score reliability and validity might be affected by the level of training and experience of the administrator. Moreover, given the length of the test and the often limited amount of time available for testing, users would benefit from analyses to explore whether administering a portion of the assessment would produce useful information about students' skills and areas of need. Additional guidance on how the PAL-II M scores should be combined with other information to design appropriate interventions would be useful. Finally, evidence of the extent to which appropriate use of PAL-II M leads to better decision making would be extremely useful for evaluating the validity of test scores and utility of the measure.

SUMMARY. The PAL-II M provides potentially useful information to help educators and clinicians understand the sources of students' difficulties in learning math and could be a valuable tool when used in combination with other information to design interventions. The administration and scoring guidelines are clearly written and relatively straightforward, and the technical information presented in the user's guide suggests that use of the test with students in Kindergarten through Grade 6 is probably justified, though other information should be incorporated into instructional decision making. Prospective users would benefit from additional validity evidence, particularly regarding the consequences of PAL-II M use for students' educational experiences and future achievement.

Test 2

Review of the Early **Math** Diagnostic Assessment by MARK J. GIERL, Associate Professor of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada, and XUAN TAN, Research Associate, Department of Educational Psychology, University of Alberta, Edmonton, Alberta, Canada:

DESCRIPTION. The Early **Math** Diagnostic Assessment (EMDA) is an individually administered, norm-referenced instrument that is derived from the Wechsler Individual Achievement Test-Second Edition (for reviews by Beth J. Doll and Gerald Tindal and Michelle Nutter of the WIAT-II, see 15:275). The EMDA is designed to measure select mathematics skills for students in PreKindergarten (PreK) to Grade 3. It is composed of two subtests. The first subtest is **Math** Reasoning, which measures skills such as counting, identifying shapes and patterns, and solving single- and multi-step problems related to time, money, measurement, and probability. This subtest is appropriate for students in PreK to Grade 3. The second subtest is Numerical Operations, which measures skills such as identifying and writing numbers, counting, and calculating. This subtest is appropriate for students in Kindergarten (K) to Grade 3. Results from EMDA subtests, when combined with 'teacher observation, parent information, the student's cumulative school records, and the instructional environment and history' (examiner's manual, p. 3), help evaluate progress in mathematics, identify areas that may require additional instruction, and monitor skill attainment and intervention effectiveness.

Teachers and other school professionals can administer and score the EMDA. The examiner's manual provides detailed information about the administration and scoring procedures. It also provides information on interpreting EMDA scores. Start points (different across grades) and discontinue rules (stop after six consecutive errors) are clearly stated. Administration time varies according to the grade of the examinee and the number of subtests administered. Generally, however, the test can be administered within 20 to 30 minutes. Testing instructions for examinees with special needs are also provided.

The test administrator enters the student's answers and scores into the record form. A total raw score is calculated using number correct for each subtest. This raw score can also be converted to a percentile range that, in turn, yields a 'category of achievement' and a standard score ($M = 100$, $SD = 15$). Students in PreK to K fall into three

categories of achievement—Emerging, Basic, and Proficient—whereas students in Grades 2 to 3 fall into the categories Below Basic, Basic, and Proficient. These categories apply to scores on both the **Math** Reasoning and Numerical Operations subtests. 'Emerging' students (i.e., PreK to K) and 'Below Average' students (i.e., Grades 1 to 3) have raw scores that fall between the 1st and 29th percentile; 'Basic' students have raw scores that fall between the 30th and 69th percentile; 'Proficient' students have a raw scores that falls between the 70th and 99th percentile. Unfortunately, neither the rationale for these categories of achievement nor the score range is justified or substantiated empirically (using, for example, standard setting procedures; see Cizek, 2001).

DEVELOPMENT. The EMDA originated from the **Math** Reasoning and Numerical Operations subtests of the WIAT-II. Item development started in 1996 with an expert advisory panel consisting of educators, school psychologists, university professionals, and parents. The blueprint for the constructs measured by the EMDA was developed by incorporating the 'recommendations of the advisory groups, review of state standards and curricula, and an extensive search of the literature' (examiner's manual, p. 35). The curriculum areas measured by the EMDA were based on the Principles and Standards for School Mathematics (National Council of Teachers of Mathematics, 2000), state standards, and recommendations from the expert advisory panel.

Once the item pool was developed, the EMDA was field tested. Unfortunately, the field test samples and procedures are vaguely described in the examiner's manual. Field testing was conducted in the Spring of 1997 with 100 students in K to Grade 2. Then, an additional 350 students in the 'three targeted grades' were tested in the following year. Basic information about gender and ethnicity is provided for the K to Grade 2 sample but no information is provided about the characteristics for the 350-student sample. Sixty PreK 'at risk' students were also included (these PreK students may have been in either the 100- or 350-student sample, but it is not clear from the description in the examiner's manual). Data from the pilot studies were then used for item analyses. The outcomes from these analyses, according to the examiner's manual, guided the item modification efforts and helped select the item administration order. The items retained from the field tests, based on their statistical characteristics and curriculum representativeness, were administered to the standardization sample.

TECHNICAL.

Standardization. Norms for the EMDA are based on the WIAT-II standardization sample. The standardization sample was collected between 1999 and 2001 from 1,374 school-aged children (i.e., 5 to 8 years old) from PreK to Grade 3. A stratified random sampling design was used to ensure the sample represented the 1998 Census of the United States for students in PreK to Grade 3. Grade, gender, race/ethnicity, geographic region, and parents' education level served as the stratification variables. The sample also included students from public and private schools; students with learning disabilities (including speech/language impairments, emotional disorders, mild mental impairment, attention deficit disorder, and physical impairment); and students from gifted and talented programs. However, it excluded students who could neither speak nor understand English.

To ensure the data from the standardization sample had integrity, two quality-control procedures were used. First, each primary examiner received training from the test publisher on the proper administration and scoring of the WIAT-II. Second, each protocol scored by a primary examiner was rescored by two independent examiners. The protocol scores were then compared across the examiners for consistency (the results from these comparisons are not reported in the examiner's manual).

Reliability. Reliability results for the EMDA consist of estimates for score precision, consistency, and stability. Split-half reliability, which measures score precision, ranged from .71 to .93 using data from the EMDA standardization sample. The standard error of measurement, which can be used to evaluate score consistency, ranged from 3.97 to 8.08 standard score units. Test-retest reliability, an index of score stability, ranged from .96 for **Math** Reasoning to .92 for Numerical Operations using a testing interval of 7 to 45 days (M = 10 days). It should be noted, however, that the test-retest reliability analyses were based on the data for 123 K to Grade 3 students drawn from the WIAT-II standardization sample. Although the gender and ethnicity proportions for the test-retest sample are presented, little additional information is provided that would allow the reader to compare the test-retest sample to the EMDA standardization sample or, more importantly, the 1998 census data. The mean testing interval of 10 days could also pose a threat to the accuracy of the information on score stability because this interval is relatively

Comment [L8]: Reliability: internal consistency

Comment [MG9]: Reliability: Test-Retest

short. There is no information on score stability for the PreK students.

Validity. A contemporary view on test validation is provided where content, construct, and criterion-related evidence is described and presented. Content-based evidence is presented both in the Development and Validity sections in the examiner's manual. As we described earlier, results from the expert advisory panel and field test analyses guided item selection. Further, classical and modern item analyses were conducted using data from the standardization sample to ensure that items had appropriate psychometric characteristics. Unfortunately, the results from the item analyses are only described in the examiner's manual, not presented. For example, the authors state that the item-total correlations were inspected for each test and age level, and any item with a correlation less than .20 was considered for revision and replacement. But we are not told if the discrimination index is the biserial or point-biserial correlation; we are not shown the discrimination values for the items on the final form; and we are not told how the items were revised or if items with low discrimination were replaced. Another example: The authors claim that extensive item-bias studies were conducted using data from the field test and standardization samples, and any item noted as problematic according to empirical criteria was evaluated for deletion. Again, however, we are not told which groups were compared in the item bias study, which statistical and substantive methods were used to detect and interpret bias, what criteria constitute a problematic item, or whether problematic items were modified and/or deleted. In short, few empirical results are presented that would allow the reader to independently evaluate the content-related evidence; instead, the reader must rely on descriptive summaries from the test publisher.

Comment [L10]: Validity: test content

Comment [MG11]: Validity: internal structure

A critical line of evidence to support the diagnostic value of the EMDA stems from information about the **Math Reasoning and Numerical Operations** constructs (Tate, 2002). To support the validity of the construct interpretation, the authors present results where the EMDA scores are correlated with other achievement constructs, including the other WIAT-II subtests, WIAT-II composites, and Wechsler intelligence scales (i.e., Wechsler Preschool and Primary Scale of Intelligence-Revised and Wechsler Intelligence Scale for Children-Third Edition) across five age groups (i.e., 5- to 9-year-olds). As expected, the EMDA scores were more highly correlated across all age groups with scales measuring mathematical constructs than with scales measuring nonmathematical constructs, such as reading. The correlations between the EMDA and Wechsler scales were variable, ranging from .25 to .81. Regrettably, the construct validity section does not include dimensionality results from either parametric (e.g., factor analysis) or nonparametric (e.g., DIMTEST, DETECT, or HCA/CCPROX; see Stout et al., 1996) procedures—a noteworthy omission in light of the fact that the EMDA is designed to measure two distinct constructs.

Comment [MG12]: Validity: relationships with other measures

Criterion-related evidence is documented, in part, by correlating the EMDA scales with select achievement tests. Correlations between the EMDA and the Wechsler Individual Achievement Test (WIAT) **math** scales ranged from .78 to .82; correlations between the EMDA and the Wide Range Achievement Test-Third Edition (WRAT-3) arithmetic scale ranged from .67 to .77; and the correlation between the EMDA Numerical Operations scale and Differential Ability Scales (DAS) Basic Number Skills scale was .75. However, in all three studies the sample sizes are small, ranging from 27 to 70 examinees, and little information is provided about the composition or characteristics of the samples. In fact, sample-related problems characterize all the studies presented in the criterion-related validity section of the examiner's manual. For instance, correlations between EMDA scores and teacher-assigned school grades are presented. The teacher-assigned mathematics grades correlated .35 with the EMDA Numerical Operations score and .51 with the EMDA **Math Reasoning** score. However, the student sample ranged in age from 6-18 years ($M = 14$ years) which is well beyond the age range in the EMDA standardization sample. Results from studies using special groups are also reported. The special groups included students who were gifted, mentally challenged, learning disabled, hearing impaired, or speech and/or language impaired. In almost every comparison, a matched control group performed differently than the special group in question. However, most of the special group samples, again, go beyond the age range for the EMDA standardization sample. For example, the gifted sample of 123 students ranged in age from 6 to 17 ($M = 11$) years; the mentally challenged sample of 39 students ranged in age from 7 to 14 ($M = 10$) years; the learning disabled sample of 81 students ranged in age from 7 to 18 ($M = 12$) years; and the hearing-impaired sample of 31 students ranged in age from 6-13 years, and included some 17-year-olds ($M = 11$ years). The speech- and/or language-impaired sample of 49 students ranged in age from 5-8 years ($M = 7$), which is comparable to the EMDA standardization sample, but there is no information on the sample

composition making it impossible to compare the speech- and/or language-impaired group with the EMDA standardization sample.

Comment [MG13]: Validity: Relationships with other measures

COMMENTARY AND SUMMARY. The EMDA contains two subscales—**Math** Reasoning and Numerical Operations—derived from the WIAT-II. As such, the EMDA is not a new instrument but a set of subtests from an existing instrument, the WIAT-II. The EMDA has important strengths. For example, it is easy to administer and score; it can be used with young children who range from PreK to Grade 3; it is developed to reflect the Principles and Standards for School Mathematics based on consultation with a diverse expert advisory panel; and it is anchored to norms that represent the 1998 American Census for students in PreK to Grade 3. However, the EMDA also has important weaknesses stemming, largely, from technical limitations and inadequacies. For instance, the 'categories of achievement' (examiner's manual, p. 23) that are used to promote score interpretation are neither justified nor substantiated empirically; the samples used initially to develop the EMDA items and then validate the EMDA scales are relatively small and poorly described; and, most importantly, the evidence to support the **Math** Reasoning and Numerical Operations constructs is incomplete, as no dimensionality analyses are reported. These weaknesses should be addressed through additional validation studies designed to provide users with more information about the technical characteristics and quality of this test.

REVIEWERS' REFERENCES

Cizek, G. J. (2001). Setting performance standards: Concepts, methods, and perspectives. Mahwah, NJ: Erlbaum.

National Council of Teachers of Mathematics. (2000). Principles and standards for school mathematics. Reston, VA: Author.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement*, 20, 331-354.

Tate, R. (2002). Test dimensionality. In J. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 181-211). Mahwah, NJ: Erlbaum.