Technical Properties of PAL-II

Development and Standardization of PAL-II ...................... 2

Research Guidelines ................................................................. 2
Major Research Phases ............................................................... 2
  Pilot Phase ............................................................................... 2
  Tryout Phase ........................................................................... 2
  Standardization Phase ............................................................. 3
  Final Production and Evaluation .............................................. 3
Quality-Control Procedures ...................................................... 3
  Qualifying Examiners ............................................................. 3
  Quality Assurance of Scoring and Data Entry ......................... 4
  Other Quality Assurance Procedures ....................................... 4
Standardization and Norms Development ............................... 4
  Locating and Testing the Sample ............................................ 4
  Description of the Standardization Sample ........................... 5
  Representativeness of the Sample ........................................... 6
Derivation and Interpretation of Scores .................................... 6
  Raw Scores ............................................................................. 6
  Determination of Start Points, Discontinue Rules, and Stop Points ........................................ 6
  Derivation of Scaled Scores .................................................... 7
  Derivation of Cumulative Percentages ................................... 7
  Derivation of Base Rates ....................................................... 7
  Derivation of Fluency Scores .................................................. 7
  Derivation of RAN/RAS Rate Change Scores ......................... 8
  Derivation of Composite Scores ............................................ 8
Evidence of Reliability .............................................................. 2
  Reliability in Diagnostic and Neuropsychological Assessments ............................................ 2
  Test Reliability ...................................................................... 3
  Evidence of Internal Consistency and Overall Reliability Statistics for Scaled Scores .......... 3
    Standard Errors of Measurement and Confidence Intervals ................................................. 4
  Evidence of Test-Retest Stability ............................................ 5
  Evidence of Inter-Scorer Agreement ....................................... 6
Validity .................................................................................... 3
  Evidence of Content Validity .................................................. 3
    Evidence Based on Response Processes .................................. 3
  Evidence of Construct Validity ................................................. 4
    Intercorrelations of Subtest Scores ......................................... 4
  Evidence of Concurrent Validity .............................................. 6
    Relationships to Other Measures .......................................... 6
  Evidence Based on Consequences of Testing .......................... 10
Development and Standardization of PAL–II

Research Guidelines

The Standards for Educational and Psychological Testing (hereafter referred to as the Standards; American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) guided the process of developing the PAL–II and provided criteria for the “evaluation of tests, testing practices, and the effects of test use” (p. 2). The reader is referred to the Standards for a comprehensive discussion of these and other issues related to test development and evaluation.

Major Research Phases

The PAL–II Diagnostic Assessment for Reading and Writing (PAL–II RW) and the PAL–II Diagnostic Assessment for Math (PAL–II M) were developed in three phases: a pilot (2000); a tryout (2004–2005); and a national standardization and validation phase (2006).

Pilot Phase

A small pilot study evaluated the item content, subtest administration directions, and scoring rules for the PAL–II M subtests. Based on results of these pilot studies the content of the subtests was modified for instructions and grade appropriateness. In addition, the pilot phase determined that an increase in content coverage was needed. This resulted in the creation of additional subtests and modification of the subtests used in the pilot phase.

Tryout Phase

New tests of Morphological Decoding and Morphology and Syntax that were based on published research were included in this phase; they were: Are They Related?, Does it Fit?, Sentence Structure, Find the True Fixes, and Morphological Decoding Fluency. Verbal Working Memory subtests, which are based on research under review, were included in the tryout also; they were: Letters, Words, and Sentences: Listening and Sentences: Writing. The sensorimotor subtests, Finger Sense and Oral Motor Planning, and RAN/RAS subtests were also developed for the tryout edition of the PAL–II RW. Writing composition subtests were also developed, based on published research, and tested during this tryout.

The PAL–II M subtests included in the tryout edition were Numeral Writing, Fact Retrieval, Numeric Coding, Place Value, Finding the Bug, and Multi-Step Problem Solving. The administration time for tryout depended upon the level tested. There were two grade bands tested, Level 1 (K–2) and Level 2 (3–6). The average time taken for the entire sample to complete every subtest ranged between 3 to 5 hours. On average, the administration time ranged from 1 hour (Kindergarten) to 3.5 hours (Grade 6).

Examiners were recruited throughout the Summer of 2004. Children were recruited throughout Summer, Fall, and Winter of 2004. The ship date for the tryout materials to examiners was late August, 2004. The examiners were required to complete a review case before formal testing of children. Formal tryout testing finished in February of 2005. The sample included 369 normally developing children in Kindergarten through Grade 6 with approximately 50 children tested from each grade. The sample was stratified by sex, ethnicity, and region. An additional 31 children diagnosed with a learning disability or ADHD were included in the sample.
Based on the results of the tryout, changes were made to subtest content, administration, and scoring. These were done to improve the reliability and ease of use of the subtests. Examiner feedback was solicited and their recommendations considered in the development of the standardization materials.

A mini-pilot of 91 cases was completed to evaluate the possibility of developing a measure of Part-Whole Relationships. The instructions and items were evaluated for appropriateness for Grades 2–6. The mini-pilot indicated that the test had good psychometric properties. Items were then modified and selected for further development in the standardization edition.

**Standardization Phase**
In preparing the standardization version of the PAL–II RW and PAL–II M, final modifications were made to the subtests, and grade levels were determined for each subtest. The standardization and validation phase was conducted from Spring to Fall of 2006. The standardization version of the PAL–II RW and PAL–II M was administered to 700 children between K–Grade 6. The standardization sample was stratified by age, race/ethnicity, sex, parent education, and geographical region. During standardization, two subtests were modified to improve their psychometric properties. The subtests were Fact Retrieval and Computational Operations. The changes in administration and scoring yielded very good psychometric qualities. These subtests were normed on a smaller stratified sample of approximately 50 cases per grade level.

The entire grade-appropriate battery was administered to all children in the standardization sample. Administration times for the PAL–II Reading and Writing and Math subtests during standardization were: K took 1.5 to 3 hours, Grades 1–2 took 1.5 to 3.5 hours, and Grades 3–6 took 3 to 7 hours. Examiners administered the test on up to 3 separate days. Final versions of the subtests should require less testing time.

Some of the children completed additional validity tests including the DAS–II, the WNV, and the NEPSY–II. The validity tests were administered on separate days from the PAL–II RW and PAL–II M batteries.

**Final Production and Evaluation**
After reviewing the standardization and validation data, the final selection of subtests and scoring procedures was made. Scoring and administration rules were finalized for item order and start, stop, and discontinue rules. Scoring studies were completed for writing subtests, and the final scoring procedures developed for all subtests.

**Quality-Control Procedures**
Specific procedures were utilized during the development of the PAL–II to ensure the quality of the data collected and to assist in the development of the final scoring criteria. The quality-control procedures used in collecting the PAL–II standardization data were designed to facilitate proper test administration and to ensure that test responses were accurately scored.

**Qualifying Examiners**
One of the first steps taken was to recruit examiners with experience testing children of various ages. Potential examiners completed a questionnaire by supplying information about their educational and professional experience, administration experience with academic achievement tests and other cognitive measures, certification, and licensing status. Those selected as potential standardization examiners were familiar with assessment practices. Most of the examiners were certified or licensed professionals working in private or public facilities. Potential standardization examiners were provided training materials, which consisted of a training video, a summary of common administration, recording, and scoring errors, and a training quiz. Examiners were selected based on their performance on the training quiz. Any errors or omissions were reviewed.
with the examiner. Examiners were then required to submit a review case prior to testing additional children. Every attempt was made to discuss administration and scoring errors on the review case within 48 hours of receipt, and any errors in administration, recording, or scoring were discussed with the examiner. Omitted subtests were collected during a follow-up session with the child. Subsequent cases were reviewed within 72 hours of receipt, when possible, and all errors resulting in loss of data were discussed with the examiner. The majority of the examiners submitted protocols that were correctly administered and scored. To address the few frequently occurring administration and scoring problems, periodic newsletters were sent to all examiners alerting them to potentially problematic areas.

Quality Assurance of Scoring and Data Entry
All scorers had a minimum of a bachelor’s degree and attended a two-week training program led by members of the development team. All scorers received ongoing feedback on scoring errors and were given additional training as needed. On more complex writing tasks, many of the scorers were experienced specifically at scoring writing measures and many had advanced degrees.

Each protocol collected during the pilot, tryout, and standardization phases of development was scored and entered into a database by two different qualified scorers working independently. Any discrepancies between the two scorers were resolved daily by a third highly trained scorer (resolver) or a member of the development team. The resolvers were selected based on their previous scoring experience, demonstration of exceptional scoring accuracy, and ability to identify problems that required elevation to the development team.

To prevent scorer drift, scorers did not discuss scoring rules with other scorers and directed any scoring questions to the resolver or a member of the development team. Members of the development team second scored or resolved cases throughout data collection to ensure scorers were adhering to the scoring guidelines. Scorers received feedback immediately to prevent repetition of the errors and to correct for scoring drift.

Other Quality Assurance Procedures
In addition to scoring and data entry quality assurance procedures, several other procedures were employed to ensure the consistency of data handling. A computer program automatically checked the values entered by scorers for out of range values. For example, the program would not accept values that fell outside of a specified range. After all protocols were double scored and resolved, a data clean-up team performed additional checks on the data to address unusual and extreme values or contradictions in scoring information. Members of the development team randomly checked and rescored multiple protocols and compared the results to the final data file.

Standardization and Norms Development

Locating and Testing the Sample
Approximately 225 examiners across the United States were selected to participate in the PAL–II standardization and validation. Examiners were selected based on the populations they had access to in regard to age, socioeconomic status, race/ethnic composition, and their experience with assessing children with individually administered tests.

Parent consent forms were mailed to each examiner. The consent form requested the child’s age, sex, race/ethnicity, and education level of the parent(s). A sampling matrix database was prepared based on the stratification variables of the project and the census percentages for each variable. Children were randomly selected to participate in the project if they met the demographic qualifications of the matrix database. The following list cites the exclusionary criteria for the standardization sample of the PAL–II.
• Diagnosed impairment and/or receiving services for delay in cognitive, academic, motor, language, social-emotional, or adaptive functioning
• Referred for assessment because of a suspected delay (including language delay)
• Diagnosed with and/or receiving services for one or more of the following:
  ◦ ADHD and or Learning Disability (including Nonverbal Learning Disability and/or Asperger’s)
  ◦ Expressive and/or Receptive Language Disorder
  ◦ CAPD or Limited English proficiency
  ◦ Psychiatric Disorder (e.g., anxiety, depression, OCD, bipolar)
  ◦ Chromosomal abnormality and/or Congenital infections
  ◦ Disorder secondary to exposure to toxic substances (e.g., fetal alcohol syndrome, lead)
  ◦ Disorders reflecting disturbance of the development of the nervous system (e.g., epilepsy)
  ◦ Acquired CNS disorders, such as brain tumor or TBI
  ◦ Intellectual Disability
  ◦ Severe Attachment Disorder
  ◦ Severe sensory impairment (including vision or hearing)
• Currently admitted to a hospital, mental, or psychiatric facility
• Currently taking prescribed medication that might depress test performance (e.g., anti-convulsants)

Description of the Standardization Sample
The PAL–II normative data presented in this manual are derived from a standardization sample that was representative of the U.S. population of children in K–Grade 6. A stratified, random sampling plan was used to ensure that representative proportions of children from each demographic group would be included in the standardization sample. An analysis of data gathered in October 2003 by the U.S. Bureau of the Census provided the basis for stratification along the following variables: age, sex, race/ethnicity, geographic region, and parent education (parent refers to parent[s] or guardian[s]). The following sections present the characteristics of the PAL–II standardization sample.

Grade
The standardization sample of 700 cases included 100 children in each of seven grade groups ranging from Kindergarten through Grade 6. In RAN–Words and RAS a smaller stratified sample of first graders (N = 62) was tested; and in Numeral Coding a smaller stratified sample of kindergarteners (N = 70) was tested.

Sex
The sample included 50 males and 50 females at each grade level.

Race/Ethnicity
For each age group in the standardization sample, the proportions of Whites, African Americans, Hispanics, and other race/ethnic groups were based on the race/ethnic-group proportions of children in K–Grade 6 in the U.S. population according to the October 2003 census survey. Each
child in the standardization sample was categorized by his or her parent(s) as belonging to one of the following race/ethnic groups: White, African American, Hispanic, or Other. Native American, Eskimo, Aleut, Asian, and Pacific Islander categories were included in the Other category.

**Geographic Region**

The United States was divided into the four major geographic regions specified in the Census Bureau reports: Northeast, North Central, South, and West. Children were selected for the normative group in accordance with the proportions of children living in each region.

**Parent Education**

The sample was stratified according to the following three parent education categories:

- 0–11 years
- 12 years (high school graduate or equivalent)
- Greater than 12 years (some college or Associate’s degree, college degree, graduate degree)

Information on parent education level was obtained from parents’ responses to a question that asked them to specify the highest grade completed by each parent living in the home. For children residing with both parents, the average of the two education levels was used.

**Representativeness of the Sample**

Tables P.1, P.2 and P.3 summarize the distribution of parent education, race/ethnicity, and geographic region in the total PAL–II standardization sample. Tables P.1 to P.3 present detailed demographic characteristics of the standardization sample. These data show that for the stratification variables selected, the PAL–II standardization sample closely approximated the census data.

**Derivation and Interpretation of Scores**

**Raw Scores**

**Scoring Studies**

Several scoring studies were conducted in order to further refine the scoring criteria and increase scoring accuracy. In the PAL–II RW battery, the following subtests went through extensive scoring evaluation: Alphabet Writing, Copy Task A and B, Verbal Working Memory Sentences: Listening and Sentences: Writing, Narrative Composing, Expository Note Taking, and Expository Report Writing. On the PAL–II M battery, the subtests requiring extensive scoring included Numeral Writing, Spatial Working Memory, and Part-Whole Time. All cases were scored by trained scorers and scoring consistency was monitored throughout the scoring process.

**Determination of Start Points, Discontinue Rules, and Stop Points**

Because the PAL–II is designed for K–Grades 6 some items are too difficult for younger children, and other items are too easy for older children. To avoid frustrating the child by administering overly easy or overly difficult items, start points, discontinue rules, and stop points are used in subtests assessing multiple ages. In the standardization version of the PAL–II, subtest items were ordered according to increasing difficulty as indicated by pilot and tryout phase data. Start points, discontinue rules, and stop points for the subtests were set generously in order to allow the child to attempt all of the items that he or she might be expected to pass, yet limit the number of items presented. Final adjustments in the rules for starting and stopping each subtest were made on the basis of empirical studies of the standardization data. If 95% or more of the children in a
particular age group passed the first several items administered, the start point for the age group was moved to a later item.

To set the final discontinue rule for each subtest, the probability of obtaining additional points after each of several possible discontinue points was examined. For example, after four consecutive incorrect responses on the Numeric Coding subtest were given, the probability that any child would pass additional items on the subtest became very small. For this reason the final discontinue rule was changed from five incorrect responses in the standardization version to four in the final version. Stop points were determined in a similar manner and placed at a point beyond which less than 5% of children of a certain age earned credit.

**Derivation of Scaled Scores**

The PAL–II normative information was developed using the method of inferential norming (Wilkins, Rolfhus, Weiss, & Zhu, 2005). Various moments (means, standard deviations, and skewness) of each score were calculated for each of the seven grade levels of the PAL–II standardization sample. The moments were plotted across grade, and various polynomial regressions, ranging from linear to 4th degree polynomials, were fit to the moment data. Functions for each score moment were selected based on consistency with underlying theoretical expectations and the pattern of growth curves observed in the normative sample. For each subtest, the functions were used to derive estimates of the mid-year population moments. The estimated moments were then used to generate theoretical distributions for each of the reported normative age groups, yielding percentiles for each raw score. For subtests where theoretical distributions could not be generated due to the properties of the estimated moments, the sample distributions were adjusted according to the selected polynomial function to achieve the mid-year percentiles for the scores. These percentiles were converted to standard scores with a mean of 10, a standard deviation of 3, and a range of 1–19. The progression of standard scores within and across age groups was then examined, and minor irregularities eliminated by smoothing.

**Derivation of Cumulative Percentages**

Cumulative percentages are reported for scores that have highly skewed underlying distributions. The percentages of cases were classified by percentage bands: \( \leq 5, 6–10, 11–25, 26–50, > 50 \). The progression of scores within and across grades was examined, and minor sampling irregularities were eliminated by smoothing.

For the cumulative percentages, the ability to identify a child’s performance as functioning below normal was determined to be crucial. Thus, the relative performance of the child is seen as more useful than the exact percentage obtained. This is especially true in subtests for which cumulative percentages are reported because the distributions were found to be non-normal and identification of a specific percentile rank could be misleading. Therefore, the range is truncated and primarily focused on differentiating low levels of performance rather than above average or superior performance.

**Derivation of Base Rates**

Base rate tables are provided for scores that are categorical rather than ordinal. The base rate is the percentage of children found in a specific category. These base rate scores refer specifically to several observation rating forms for quality of writing. If a particular rating occurs commonly within a specific grade, it is unlikely to be clinically meaningful. If a specific category is rarely rated in a particular grade, then the rating may be clinically meaningful. These scores may be used to track improvement in specific qualitative aspects of compositional writing (e.g., use of punctuation, legibility).

**Derivation of Fluency Scores**

Several subtests allow the examiner to calculate a fluency score. These scores were developed to integrate speed of performance with accuracy of performance. The concept behind the
methodology is that fast performance alone does not indicate fluency; rather, fluency is fast, accurate performance. Within levels of performance, some children are faster and more fluent, which is to say there are slow children at all levels of accuracy and fast children at each level of accuracy. The accuracy scores are segmented into ability bands, and within each ability band, total time scores were plotted. The faster children within the ability band receive a higher adjusted score than the slower children. Children in the higher accuracy groups receive higher adjusted scores than those in the lower accuracy groups, with some overlap in adjacent groups (e.g., very fast performers in the next lower group score similar to the very slow performers in the next higher group). The adjusted score is then normed by grade to account for differences in accuracy scaled-score and time distributions within grade. High scaled scores indicate fast and accurate performance while low scaled scores indicate slow and inaccurate performance.

Derivation of RAN/RAS Rate Change Scores
The rate change scores on the RAN/RAS subtests were devised to be an indicator of consistency of RAN/RAS speed and an estimate of slowing or speeding up of performance from the beginning to the end of the specific condition (e.g., letters, words). Comparing two timed scores requires that the initial speed of performance is taken into account because final speed and initial speed comparisons are not independent. For instance, if a child were to complete row 1 in 4 seconds, there is very little room to show faster performance and a small increase in speed may be meaningful because the initial rate was very fast. If another child completes the first row in 20 seconds, he or she has a lot of room for improvement.

A one-second change (i.e., faster time) in a child starting out with a first row time of 4 seconds is a 25% improvement in his or her performance, while a one-second improvement in a child with a first row time of 20 seconds is only showing a 5% speed improvement. The conversion tables for the rate change scores take into account the child’s initial speed and differentially weight improvement based on row 1 time.

It is important to keep in mind that you are comparing the time it took to complete row 1 to the time it took to complete row 4 (i.e., do not use the cumulative time that is marked in the row box in the Record Form, but subtract the total cumulative time at row 4 from the total cumulative time at row 3). After obtaining the time for row 1 and row 4, subtract row 4 from row 1 to determine if performance was faster or slower. Please refer to the Administration and Scoring Manual for a discussion of how to convert the raw total score to the scaled score. High scores indicate that the child was getting relatively faster or maintaining an initially very fast performance rate over the course of the RAN/RAS task. Low scores indicate that a child is getting slower or maintaining a very slow initial speed of performance over the course of the RAN/RAS tasks.

Derivation of Composite Scores
The PAL–II RW and PAL–II M batteries yield composite scores for domains that have more than a single subtest. Composite scores are derived by summing the contributing subtest scaled scores. The distribution of the sum of scaled scores was used to derive corresponding mid-interval percentiles that were converted to composite scores. The resulting composite score distributions were smoothed visually to remove any irregularities and to ensure that the distribution was approximately normal, while attempting to keep the means and standard deviations. Using subtest scaled scores is preferable to summing the total raw scores of each individual subtest. By using scaled scores, each subtest is equally weighted rather than the subtest with the largest number of data points or biggest variance being weighted more heavily than tests with fewer items.

The composite scores will have better psychometric properties (e.g., reliability) than the individual subtests themselves and provide a stable estimate of the overall domain. As with any composite score, the examiner is encouraged to not report the composite when the subtests scaled scores are significantly different from one another. Unlike composite scores on other tests (e.g., DAS–II, WIAT–II), the PAL–II uses a metric with a mean of 10 and a standard deviation of 3
rather than a mean of 100 and a standard deviation of 15. The composite scores on the PAL–II batteries often summarize subcomponent measures of a single test (e.g., Computational Operations Part A + Part B + Part C). In addition, PAL–II subtests measure sub-processes (e.g., phonological coding rather than reading comprehension) of academic or cognitive skills. This is unlike several other test batteries where composite scores often represent higher level indexes that combine a number of related but independent procedures, or represent factorially derived higher order constructs. Also, the scaled score metric facilitates comparison between PAL–II composite scores and PAL–II subtests that do not have composites.
Technical Properties of PAL-II

Evidence of Reliability ........................................ 2
- Reliability in Diagnostic and Neuropsychological Assessments .................................................. 2
- Test Reliability .................................................................................................................................. 3
- Evidence of Internal Consistency and Overall Reliability Statistics for Scaled Scores .................. 3
  Standard Errors of Measurement and Confidence Intervals .............................................................. 4
- Evidence of Test-Retest Stability ....................................................................................................... 5
- Evidence of Inter-Scorer Agreement .................................................................................................. 6
Evidence of Reliability

The information provided here describes the psychometric properties of the PAL–II Reading and Writing (PAL–II RW) and PAL–II Math (PAL–II M) standard scores and cumulative percentages, including internal consistency, standard error of measurement, test-retest stability, and inter-rater reliability. The statistical properties of the PAL–II are used in establishing the degree of confidence with which an interpretation of performance on individual subtests may be made.

Reliability in Diagnostic and Neuropsychological Assessments

The PAL–II is a battery of subtests designed to measure various domains of reading, writing, math, and related processes in children K–Grade 6. The subtests employ a variety of stimulus-response procedures. For some subtests, the child is shown individual items of similar content but of increasing difficulty until the child meets a discontinuation criteria. On some subtests, the child tries to complete all the items as quickly as possible, and while on others the child responds to a single prompt. The tasks vary in the degree to which internal consistency measures may be applied, and in many cases test-retest procedures are the best estimate of reliability.

In order to obtain relatively pure measures of the construct being measured, a particular subtest may display range restrictions either at the upper or lower end of the distribution. For example, alphabet writing or numeral writing requires the child to write the 26 letters of the alphabet (or numbers) as quickly as possible within 3 minutes. The derived scores that analyze performance at 15 seconds and within 180 seconds yield very different distributions. In kindergartners, this is a challenging task and many children do poorly at the 15-second mark, but overall, do better within the 180-second time frame. Children in older grades perform almost perfectly on the total score but show more variance at 15 seconds. The impact is that for some measures there is more range restriction. This is important because it has a direct impact on the estimated reliability. For example, a low test-retest score may be observed in a variable that is performed almost perfectly by most normally developing children, yet it is still very useful for clinical assessment and monitoring. Other constructs, such as motor and oromotor skills will have similarly skewed distributions and may appear to have lower than expected reliabilities in children without developmental difficulties.

Estimating reliability on error measures provides even greater challenges. In typically developing children, errors tend to be minimal and highly skewed, with most children having only a few or no errors. Error scores are often associated with timed tests in which there is a trade off between speed and accuracy. Errors occur because a child is going too quickly or a child goes slower to ensure accurate performance. On reevaluation, children may change their strategy because they have already been exposed to the task and know what to expect. Some children who went very quickly and had a high rate of error will intentionally slow down and provide a more accurate but slowed performance. Other children may speed up because they feel they mastered the task on the previous assessment only to find themselves making a couple more errors. Shifts in performance such as these will attenuate the association from time 1 to time 2 but may not be indicative of how children with diagnostic conditions would approach the task. Subsequently, reliability measures in normally developing children provide only an estimate of how the test will function in a clinical setting.
Test Reliability

Test reliability is an indication of the degree with which a test provides a precise and stable measure of the underlying construct it is intended to measure. In classical test theory, any test score is an approximation of the child’s hypothetical true score, that is, the score the individual would receive if the measuring instrument were perfectly accurate. The difference between the person’s true score, which can never be known, and the score the person actually obtains on a single administration is measurement error. The smaller the average amount of measurement error on a test, the more accurately the observed score will estimate the true score. A reliable test will have relatively small measurement error and produce consistent results across administrations. The reliability of a test refers to the accuracy, consistency, and stability of test scores across situations (Anastasi & Urbina, 1997). The reliability of a test should be considered in the interpretation of obtained scores on one occasion and differences between scores obtained on multiple occasions.

Multiple procedures exist to assess the reliability of a test, yielding a variety of data regarding the test’s internal consistency, and score stability over time. The estimate of a test’s reliability is used in obtaining the standard error of measurement, which is required to calculate confidence intervals around the primary scores. The reliability procedures applied to the PAL–II RW and PAL–II M vary among the subtests based on the properties of the subtest and whether it is suited to internal consistency and stability statistics. Reliability measures are provided for all of the PAL–II RW and PAL–II M scaled scores and cumulative percentages. This section also presents results of studies that establish the level of agreement among raters on PAL–II subtests that require interpretive scoring.

Evidence of Internal Consistency and Overall Reliability Statistics for Scaled Scores

Reliability coefficients were obtained utilizing the split-half and alpha methods. In addition, stability coefficients were used on subtests for which the alpha and split-half methods were not appropriate. Internal consistency measures may be attenuated by item interdependence, especially on tasks in which children may adjust their problem-solving strategies based upon feedback received from responses to earlier items or on memory tests that have repeated trials with the same stimuli. The degree of item interdependence in a task has a bearing on the type of internal-consistency analysis that is best suited for that task.

The split-half reliability coefficient of a subtest is the correlation between the total scores of the two half-tests, corrected for the full test, using the Spearman-Brown formula (Crocker & Algina, 1986; Li, Rosenthal, & Rubin, 1996). The internal consistency reliability coefficients were calculated with the formula recommended by Nunnally and Bernstein (1994). The average reliability coefficients were calculated using Fisher’s $z$ transformation (Silver & Dunlap, 1987; Strube, 1988). Internal consistencies were provided by grade.

For measures that do not lend themselves to internal consistency procedures, test-retest stability is reported as the primary indicator of subtest/score reliability. These subtests are based primarily on speeded performance or those in which a single prompt was used to elicit the child’s response. The stability coefficients are based on the scores of children in two grade bands (i.e., K–3 and 4–6) who participated in the test-retest study. The stability coefficients used as reliability estimates are the correlation of scores on the first and second testing.

Table R.1 provides reliability coefficients for all subtest scaled and composite scores by grade and for the averages across the seven grade levels for the PAL–II RW. The results of the reliability studies indicate that most of the PAL–II subtests have adequate to high internal consistency or stability. Composite scores have the highest reliability. Among individual subtest
measures, Pseudoword Decoding, Rhyming, Are They Related, Find the True Fixes, Morphological Decoding Fluency, Compositional Fluency, and Verbal Working Memory Letters, Words, and Sentences: Listening had the highest internal consistencies. Overall the lowest reliability coefficients were related to basic academic processes calculated with test-retest reliability, such as alphabet writing and copying.

Table R.2 provides reliability coefficients for the PAL–II M subtest scaled and composite scores. The math subtests show very good reliability overall with many reliabilities over .9 and most over .8. The tests with the lowest reliability are related to more basic processes, such as counting and numeral writing and those related to sensorimotor processing.

On both batteries, some measures show a drop in reliability for a specific grade. These can be due to natural fluctuations in sampling and are random effects. In other case, there may be a ceiling issue such that a full range of scores is not obtained within that grade level. In these cases, it is important to keep in mind that superior or above average performance is not relevant to the clinical question, rather the presence and degree of impairment is essential.

**Standard Errors of Measurement and Confidence Intervals**

An individual’s obtained score on a particular instrument represents an estimate of that person’s true score. The true score is the hypothetical level of performance on the test that would be attained by that person on a test that had no measurement error. Although an individual’s true score can never be determined, statistical methods have been developed to aid the examiner in determining the degree that measurement error is present in a particular score. The reliability coefficient provides one estimate of the precision of a test. However, knowledge of the reliability coefficient provides the examiner with only a rough estimate of the degree that measurement error may be affecting a particular obtained value. The standard error of measurement (SEM) provides another estimate of the amount of error in an observed score. The SEM has an inverse relationship to the reliability coefficient, thus, the greater the reliability of a test, the lower the SEM and the greater the confidence the user can have in the precision of the observed test score. Measurement error is commonly expressed in terms of the standard deviation units, that is, the SEM is the standard deviation of the measurement error distribution. The SEM is calculated with the formula:

\[ SEM = SD \sqrt{1 - r_{xx}} \]

where SEM represents the standard error of measurement, SD is the theoretical standard deviation of the unit of the scale, and \( r_{xx} \) is the reliability coefficient of the scale. The SEM for the subtest and composite scaled scores is based on a standard deviation of 3. Table R.3 displays the SEM for all PAL–II Reading and Writing scaled scores and Table R.4 presents SEM for PAL–II Math Scaled Scores. Standard error of measurement is not reported for scores represented as cumulative percentages of base rates.

The SEM can be used to calculate the confidence interval, or the band of scores, within which the individual’s true score falls. Confidence intervals provide another means of expressing the precision of test scores and serve as a reminder that measurement error is inherent in all test scores, and that the observed test score is only an estimate of an individual’s true ability.

The standard error of measurement can be used to create confidence intervals around an observed score. The interval represents the range of scores the individual’s true score is likely to fall. The confidence interval is another means of expressing the precision of an individual test score. The achieved score on the test is only an estimation of their true ability on the construct the test purports to measure. Confidence intervals based on the standard error of measurement may be calculated using the following formula:

\[ % CI = \text{observed score} \pm Z_p(SEM) \]
where $p$ is the confidence interval (e.g., 90% or 95%) and $Z_p$ is the Z value (e.g., 1.64 or 1.96) associated with the confidence level, which may be obtained in normal probability tables. For example, a child in Grade 6 obtains a scaled score of 8 on the Part-Whole Concepts subtest. The 90% confidence interval would be calculated as (using $SEM$ from Table R.4): 90% C.I. = $8 \pm (1.64 \times .73)$ or $8 \pm 1.20$.

If a 3rd grader obtained a scaled score of 6 on Find the True Fixes, their 95% confidence interval (see Table R.3 for $SEM$) would be calculated as 95% C.I. = $6 \pm (1.96 \times 1.20)$ or $6 \pm 2.35$.

Evidence of Test-Retest Stability

To determine the stability of performance on the PAL–II scores, a sample of 129 children was given the PAL–II on two occasions. The test-retest interval ranged from 2 days to 34 days with a mean of 15 days.

The sample was divided into the following two grade groups: Grades K–3 ($n=81$) and Grades 4–6 ($n=48$). Demographic information regarding the test-retest sample is presented in Table R.5.

Means and standard deviations for scaled scores for the first and second testing are presented in Table R.6. Stability coefficients between scores obtained at the first test session and scores obtained at the second test session were derived using the Pearson’s product-moment correlation. The table reports the standard differences (i.e., effect sizes) between the first and second testing. The standard difference was calculated using the mean score difference between two testings divided by the square root of the pooled variance, computed using Cohen’s (1996) formula.

Most PAL–II reading and writing scaled scores demonstrate good test-retest reliability. Few subtests have lower than expected test-retest. The lower test-retest correlations are found primarily on handwriting measures in the older group. Range restriction is a significant issue for these measures. Sentence structure shows lower than expected test-retest correlations, indicating that changes in the score over time occur frequently within healthy controls. In general, performance increases between time 1 and time 2 assessment. The degree of change is quite small for most measures, which indicates minimal practice effects over time. Not surprisingly, subtests requiring memory performance (i.e., Receptive and Expressive Coding and verbal working memory measures) and processing speed measures (i.e., RAN scores) showed the most improvement from time 1 to time 2.

Scores presented in cumulative percentage format have significantly non-normal score distributions, range restrictions, and a high degree of skew in normally developing children. These scores are better indicators of a difficulty or problem rather than of normal or above average cognitive functioning. These factors will directly impact (i.e., attenuate) the correlation between time 1 and time 2 assessments. Table R.7 provides means and standard deviations for the raw scores for these variables and the Pearson correlations between time 1 and time 2. For measures that had zero or near zero mean scores, the correlations were not calculated due to lack of variation in the normally developing children. The test-retest correlations are generally lower than is observed for scaled scores. In general, performance improved from time 1 to time 2, but the differences tended to be very small.

The test-retest study for the PAL–II M was separated into two studies. The first study used all subtests except Fact Retrieval and Computational Operations. These subtests were revised during the standardization requiring a separate test-retest study. Demographic characteristics for the two test-retest samples are presented in Table R.8. Table R.9 provides means, standard deviations, correlation coefficients, and standardized differences between time 1 and time 2 performances for the PAL–II M subtests. Most of the math subtest scaled scores show good test-retest reliability. The lowest reliabilities were observed for Numeral Writing, particularly in older children, Finger
Sense and Spatial Working Memory Oral Total Score. These scores have range restriction in healthy developing children, which will reduce the correlation coefficient. The composite scores for the subtests typically had the highest test-retest correlation, indicating very good stability in those scores. Level of performance was typically better on retest; however, the differences were very small with only a few tasks having a moderate effect sizes.

Table R.10 provides raw score means and standard deviations, correlation coefficients, and standard differences for the PAL–II M cumulative percentages test-retest study. For scores with 0 or near 0 means, no correlation was reported. These distributions are highly skewed and have restricted range. The data are presented here to show level of change from time 1 to time 2. In healthy developing children, errors will occur infrequently, particularly in children in Grades 4–6, and will yield low correlations over time.

Evidence of Inter-Scorer Agreement

All PAL–II protocols were double-scored by two independent scorers for most subtests, and evidence of inter-scorer agreement was obtained using all cases entered for scoring. For some subtests, the scores were single scored with a percentage of cases being double scored to insure scoring accuracy and monitor for scorer drift. For subtests where the scoring criteria are simple and objective, inter-scorer agreement was very high, ranging from .95 to .99. Although most of the PAL–II subtests are objectively scored and require minimal interpretive scoring, there are several exceptions. Alphabet Writing, Copy Tasks A and B, Numeral Writing, Narrative Composition, Expository Note-Taking, Expository Report Writing, Verbal Working Memory Sentences Writing, Part-Whole Time, and Spatial Working Memory are the subtests in the PAL–II RW and M batteries that required application of subjective scoring criteria. The Alphabet Writing and Copy A and B Tasks required scorers to determine the legibility of letters. Numeral Writing was scored for legibility of written numerals. Narrative Composition was scored for number of words, spelling, and complete sentences. Expository Note Taking was scored for the quality of the notes, and the Expository Report Writing was scored for quality of using notes to write the report and the quality of the writing. Verbal Working Memory Sentences: Writing was scored for the psychological relevance of the response to the prompts and the use of complete sentences. The Part-Whole Time subtest was scored for accuracy of the clock drawing representing specific time. Spatial Working Memory was scored for correct number of dots, the correct location, and the correct spatial pattern.

To determine the degree to which trained scorers were consistent in scoring these subtests during standardization, inter-rater reliability was calculated as percent agreement rates between trained scorers. The number of scorers assigned to specific subtests varied from 10 to 20. Scorer agreement was monitored closely and individuals struggling to apply the scoring rules correctly were dismissed. Scoring agreement rates varied by subtest and the complexity of the content being scored. Inter-scorer agreement rates for the subjective subtests were as follows: Alphabet Writing (91% to 98%); Copy Task A (88% to 96%); Copy Task B (91% to 97%); Compositional Fluency (49% to 74%); Expository Note-Taking (82% to 100%); Expository Report Writing (77% to 91%); Verbal Working Memory Written Sentences (82% to 100%); Numeral Writing (94% to 98%); Part-Whole Time (88% to 93%); and Spatial Working Memory (96% to 98%). Overall, the scorers showed good agreement. On more complex tasks, the range of agreement was larger.
Technical Properties of PAL-II

Validity ................................................................. 2

Evidence of Content Validity ................................................................. 3
Evidence Based on Response Processes ......................................................... 3
Evidence of Construct Validity ................................................................. 4
Intercorrelations of Subtest Scores ................................................................. 4
Evidence of Concurrent Validity ................................................................. 6
Relationships to Other Measures ................................................................. 6
Evidence Based on Consequences of Testing ................................................. 10
Validity

PAL–II is distinctive in that it assesses a set of processes that were first validated in a substantial amount of research by the author and other investigators before translating it into psychometric measures. Moreover, PAL–II is a three tiered assessment-intervention system that ultimately will be evaluated by its treatment validity—long term track record in improving student learning outcomes in reading, writing, and math in K–Grade 6.

Nevertheless, given the importance of measurement issues, three traditional psychometric validity studies were conducted. The purpose of the first two validity studies was to evaluate whether PAL–II assesses processes that are not fully accounted for by cognitive assessment of intellectual abilities. Unless it does, test results are redundant with intellectual ability test results and there is no justification for the additional testing time required to do assessment of processes. The purpose of the third validity study was to evaluate whether PAL–II assesses processes that are not fully accounted for by a test battery that assesses neuropsychological systems from a developmental perspective not explicitly linked to literacy learning. PAL–II is designed to be used with broadband assessments of cognitive and neurodevelopment while providing supplementary narrowband assessment of cognitive and neuropsychological processes that are more directly linked to explaining why individual students struggle with specific reading, writing, and math skills, and to design instructional interventions linked to research-supported instructional practices.

From a psychometric perspective, the validity of a test is the single most fundamental and important aspect of test development and evaluation (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999; Angoff, 1988). Three major types of validity: content, concurrent, and construct validity are critical for evaluating a test. A test has content validity if it adequately samples relevant aspects of the construct being measured. It has concurrent validity if scores are shown to be related to specified external criteria, such as performance on some other measure or group membership. A test has construct validity when the construct purported to be measured by the test is actually measured. Evidence of a test’s construct validity could come from many different sources, including expert review, multi-trait–multi-method studies, and clinical investigations.

Contemporary definitions of validity describe lines of evidence of validity as opposed to different types of validity. As stated in the Standards, “evolving conceptualizations of...validity no longer speak of different types of validity but speak instead of different lines of validity evidence, all in service of providing information relevant to a specific intended interpretation of test scores” (AERA et al., 1999). Validity is described as the degree to which evidence supports the interpretation of test scores for their intended purposes; therefore, the examination of a test’s validity requires an evaluative judgment by the test user.

Although test developers are responsible for providing initial evidence of validity, the examiner must evaluate if the evidence supports the use of the test for his or her intended purpose. Assessment of test validity is a multifaceted and long-term process that begins at the inception of the test and frequently continues throughout the life of the instrument. This does not mean that a test is valid only for the settings and purposes stated by the test developers. The examiner must determine the validity of the test for the purposes of the specific evaluation. A comprehensive evaluation of a scale’s validity evidence also includes an examination of the relevant literature on previous versions of a scale, as well as the literature that results from the use of a newly revised measure for different purposes, in different settings, or with different populations. It is expected that future use of the PAL–II will lead to an expanding base of evidence for the scale’s validity.
This section provides information regarding the validation procedures used and the studies conducted during the development of the PAL–II. This section summarizes the results of multiple research studies that address the content and construct validity of the PAL–II.

**Evidence of Content Validity**

The goal of content validation is to ensure that the items and subtests composing a test adequately sample the domain of behaviors included in the constructs that the test intends to measure. The PAL–II was designed to provide a detailed assessment of reading, writing, math, and related cognitive processes. The subtests were constructed based on decades of research findings in reading, writing, and math development and their disorders. The goal of the PAL–II is not to provide another standardized measure of academic achievement. Rather, the test is developed to understand which component processes in the reading, writing, and math domains have not been acquired by the child either through inadequate instruction or due to some developmental or learning disability.

The distinction between a standard achievement test and measures of academic processes is important in the development of the task. Although an achievement test may assess math computation skills in a single subtest, a process subtest will evaluate an individual skill or skills that contribute to performance on the achievement tests (e.g., knowing how to set up a math problem on paper, recall math facts quickly, or self-monitor one’s math problem solving for errors). The content of the PAL–II was designed to supplement and improve upon the information obtained from standardized achievement tests and not replace them. The goal of PAL–II, used in conjunction with achievement tests and measures of the five domains of development (cognitive/memory, receptive and expressive language, gross and fine motor, attention and executive function, and social emotional) was to improve educational diagnosis and links between assessment and instructional treatment. The research on which PAL–II is based shows clear relationships (e.g., predictive validity or criterion validity) between the achievement domain and the related process (i.e., reading decoding predicted by phonological coding).

**Evidence Based on Response Processes**

Information on response processes was collected to provide support that the child engages in the expected cognitive processes when responding to subtest tasks. This type of evidence may be provided by theoretical or observational sources or by psychometric analysis.

The majority of subtests included in the PAL–II have strong theoretical and experiential evidence of validity based on response processes. Extensive literature reviews, expert consultations, and empirical evaluations were conducted to provide additional evidence of validity based on response processes.

Additional evidence of validity was accumulated through empirical and qualitative examinations of response processes during the scale’s development. For example, response frequencies for multiple-choice items were examined to identify any incorrect responses that were commonly given (i.e., at a frequency greater than chance). Frequently occurring incorrect responses were examined to determine their plausibility as acceptable answers. Items and stimuli were then adjusted and piloted. Extensive observational data is used throughout the PAL–II. Examiners are encouraged to evaluate the child’s performance from the perspective of the child’s test scores, but also observation of strategies child used to perform tasks and what types of errors were made or which types of tasks gave the child the most difficulty. These observational skills are important for both diagnosis and intervention recommendations.
Evidence of Construct Validity

An examination of a scale’s internal structure “can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based” (AERA et al., 1999). The pattern of correlations between subtests provides information about the internal structure of the PAL–II and about the degree of relationship among subtests measuring similar content. On the PAL–II, it is expected that subtests measuring similar constructs (e.g., phonological coding, phonological decoding) will have moderate correlations with each other. In addition, many subtests designed to measure processes within a broad domain (e.g., compositional fluency, spelling, handwriting) will correlate in the low to moderate range with one another. The tests were developed to measure sub-skills within the domain and to provide unique variance to the prediction of abilities within the larger domain. Subtests measuring primarily sensory and motor skills, processing speed, or those basic cognitive or academic skills mastered in normally developing children at a very early grade (e.g., counting) may have low correlations with other measures due to range restriction.

Intercorrelations of Subtest Scores

Several a priori hypotheses were made regarding the inter-correlation studies. Correlations would be higher among subtests measuring similar constructs (e.g., Alphabet Writing and Task A–Sentence Copying) than correlations would be between less similar constructs (e.g., Alphabet Writing and Morphological Decoding Fluency). Subtests within the same domain (e.g., Morphological Decoding and Morphological/Syntactic Coding) would in general have moderate correlations to one another and have low to moderate correlations with subtests in other domains (e.g., Finger Succession, a measure of grapho-motor planning). While some subtests are labeled in different domains, the domains themselves may have similarities in cognitive processes. Working memory subtests would have moderate correlations with constructs in the same modality (e.g., Verbal Working Memory with other language subtests and Quantitative Working Memory with other math subtests). In normally developing children, it is common to see moderate correlations among cognitive skills, while in samples of children with learning disorders or language problems some specific dissociation may be observed. The current sample was comprised of normally developing children. Tables V.1 and V.2 provide correlations by grade for the PAL–II Diagnostic Assessment for Reading and Writing (PAL–II RW) subtests. Very high correlations suggest that the subtests are providing redundant information about the same process. Low to moderate correlations show that the subtests are providing unique information about individual differences in learners’ processes related to reading, writing, or math.

A moderate correlation was observed between the handwriting tasks Alphabet Writing and Task A–Sentence Copying. These subtests had low correlations with other subtests. Task B–Paragraph Copying was moderately correlated with Compositional Fluency productivity. Both tasks require sustained effort and production. Receptive and Expressive Coding tasks correlated .47 to .58 across grades. In the younger grades, Receptive Coding had a moderate correlation with Phonological Coding subtests, but in older grades this correlation was lower. Receptive and Expressive Coding moderately correlated with Pseudoword Decoding, Morphological Decoding Fluency, and Sentence Sense. Similarly, Phonological Coding subtests had moderate correlations with each other and also with Pseudoword Decoding and Morphological Decoding Fluency. Orthographic Coding and Phonological Coding are important skills associated with real and pseudoword reading.

In the Morphological/Syntactical domain, Sentence Structure had low correlations with all other subtests. Are They Related? and Does it Fit? had low to moderate correlations across grades. These subtests had low correlations with most other measures except for low to moderate correlations with Pseudoword Decoding and Morphological Decoding subtests. The
Morphological Decoding subtests—Find the True Fixes and Morphological Decoding Fluency—had low to moderate correlations across grades. Find the True Fixes had low to moderate correlations with Sentence Sense, and Morphological Decoding Fluency had low to moderate correlations with Word Choice, Sentence Sense, and Expository Note Taking. Morphological processing measures relate to aspects of reading comprehension, spelling, and writing. Word Choice and Sentence Sense had low to moderate correlations with Expository Report Writing.

Verbal Working Memory measures had low to moderate correlations with most Reading and Writing measures across grades. Verbal Working Memory demonstrated the highest correlation with reading and writing in all grades. Working Memory was associated with Orthographic Coding, Phonological Coding, Word Choice, Sentence Sense, and Morphological Decoding more than other measures. RAN tasks had low correlations with most measures and some moderate correlations. RAS had low to moderate correlations with Pseudoword Decoding and Compositional Fluency. Sensory and motor subtests had low correlations with all subtests.

The PAL–II RW subtests demonstrated moderate correlations within most domains while some had lower correlations. The latter indicate more diverse skills within the particular construct. Multiple PAL–II measures were found to be related to reading and writing skills, which is consistent with previous research using similar tasks. Verbal Working Memory and RAS subtests also correlated with aspects of reading and writing.

Tables V.3, V.4, and V.5 display correlations among the PAL–II Diagnostic Assessment for Math (PAL–II M) subtests. Fact Retrieval and Computation Operations are presented separately from the other math measures due to the need to revise these subtests during standardization. In the lowest grade levels, Oral Counting relates to Numeral Writing and Place Value subtests. In older children, Oral Counting and Numeral Writing are unrelated to other skills. This is not surprising in normally developing children who would have readily mastered these skills. Numeric Coding showed low correlations with most other math measures.

The Place Value subtests are highly correlated. Place Value had a moderate correlation with Part-Whole Fractions and Mixed Numbers and low correlations with other math measures. The Part-Whole Relationships tasks had low to moderate correlations with each other. Part-Whole Fractions and Mixed Numbers had a low to moderate correlation with Multi-Step Problem Solving. Finding the Bug and Multi-Step Problem Solving were moderately correlated in older grades.

Most Fact Retrieval correlations were moderate, with the exception of written math computation, which has range restriction in older grades and subsequently lower correlations with multiplication and division items. Computation Operations had moderate correlations among the three tasks in that domain. Fact Retrieval had low to moderate correlations with Computation Operations. The highest associations were between Multiplication and Division items on Fact Retrieval and Computation Operations, Task C–Problem Solution.

Quantitative and Spatial Working Memory measures had low to moderate correlations with other math measures, with the moderate correlations observed in the lower grades. RAN–Single Digits and –Double Digits had low correlations with most math measures. Fingertip Writing also had low correlation with math measures.

PAL–II M measures, in general, have higher correlations within domains compared to correlations across domains, as would be expected if each measure assesses unique processes. The math subtests measure multiple aspects of math functioning and provide unique estimates of math functioning more than measurement of overlapping or similar functions. In normally developing children, working memory measures are related to math performance in younger children, while sensory and rapid naming functions have low correlations with other PAL–II M subtests.
Evidence of Concurrent Validity

The examination of the relationship between test scores and external variables provides additional
evidence of a scale’s validity. Frequently, this evidence is provided through an examination of the
instrument’s relationship to other instruments designed to measure the same or similar constructs.
Historically, this type of evidence has been referred to as concurrent validity, convergent and
discriminant validity, predictive validity, and/or criterion-related validity. An examination of the
relationship between test scores and other related variables provides important information about
what a test measures and whether it behaves as expected when related to external variables.

The evidence of divergent and convergent validity was evaluated in a series of studies that
examines the relationship between PAL–II scores and other measures of general cognitive ability
and neuropsychological functioning.

Relationships to Other Measures

The relationships between the PAL–II and the following external measures were examined:
(WNV), and the NEPSY–II. For the most part, these tests were designed to measure different
constructs than PAL–II was, so correlations were predicted to be low, thus providing discriminant
validity. The sample size and demographic characteristics of each study are presented in Table
V.6. The percentages of sample representation by grade, sex, race/ethnicity, and parent education
level are presented.

There are a number of possible influences on the validity coefficients and score differences
between two instruments, many of which may also interact with one another. Some of these
influences are: (1) time interval between tests; (2) developmental change or learning during the
time interval; (3) structural differences between the two measures; (4) regression to the mean
from the first to second testing; (5) reliability of each measure; (6) retention from first
administration; (7) differential procedural learning and practice; (8) motivation level of the child;
and (9) the Flynn effect (Flynn, 1984, 1987). These influences and their interactions with each
other should be considered when interpreting the results of the comparison studies reported in this
section (Bracken, 1992; Zhu & Tulsky, 2000). Many of these issues are discussed in detail in
Bracken (1992) or Strauss, Spreen, and Hunter (2000). Methods have been proposed to estimate
the relative contribution of these factors (McArdle & Woodcock, 1997). In addition, the tests
utilized in these concurrent validity studies are all normed based on age while the PAL–II is
normed by grade level. Within grade level, children may vary in age by almost a year. This may
place them in different age norm bands on the external criterion and the relationship between the
measures may not be as large as when two tests using the same norming method are used.

Correlation of the PAL–II With Measures of General Cognitive Ability

Performance on the PAL–II was compared to cognitive functioning as measured by the DAS–II,
and the WNV in samples of normally developing children. These studies provide an estimate of
the degree to which general cognitive ability relates to specific neuropsychological functioning in
normally developing children.

When evaluating relationships between measures of general cognitive functioning and specific
neuropsychological and academic functions, it is important to consider that general cognitive
ability measures comprise multiple, integrated cognitive functions. Verbal measures often require
intact receptive-expressive language, working memory, conceptual reasoning, and auditory
attention skills. Therefore, the correlation between a neuropsychological or cognitive process
measure and general ability may represent an association not with general cognitive ability as
much as an association with the underlying component processes. It would not be surprising to
find moderate correlations between PAL–II working memory measures and general ability scores.
Because both tests require working memory, the correlation does not represent the degree to
which general cognitive ability affects the PAL–II subtest, but it reflects the degree to which component cognitive processes required in each test are related (Weiss, Saklofske, Prifitera, & Holdnack, 2006).

**Differential Abilities Scales–Second Edition**

The DAS–II is a battery of subtests designed to measure a variety of separate and distinct areas of cognitive functioning. It includes a General Conceptual Ability composite, focused on Verbal reasoning and conceptual reasoning abilities, as well as a Special Nonverbal composite comprising subtests with no verbal components. The DAS–II subtests are further grouped into lower-level composite scores called cluster scores. The PAL–II and DAS–II were administered to 90 children in K–Grade 6. Detailed demographic information is provided in Table V.6. The instruments were administered with a testing interval of 179–550 days and a mean test interval of 355 days. Table V.7 presents the correlations, means, and standard deviations for the PAL–II Reading and Writing subtests and the DAS–II clusters and composites.

The Verbal cluster of the DAS–II had moderate correlations with PAL–II Orthographic Coding, Phonological Coding, Morphological/Syntactic Coding, Morphological Decoding, and Verbal Working Memory composites. Expository Note Taking was also related to verbal ability. The largest correlation was observed with Are They Related? (.57). Verbal cognitive functioning relates to multiple PAL–II subtests, suggesting that PAL–II assesses processes related to verbal learning.

The DAS–II Nonverbal cluster was moderately related to several subtests including: Cross-Genre Compositional and Expository Writing Total Complete Sentences (.48); Expressive Coding (.41); Are They Related? (.41); Working Memory, Words (.40); and RAS Rate Change Score (.40). The DAS–II Spatial cluster showed a nearly identical association with the same subtests with the exception that the RAS Rate Change Score was not as high. In general, PAL–II RW is only modestly associated with nonverbal and spatial reasoning abilities. The DAS–II School Readiness Index was associated with the PAL–II Phonological Coding Composite and Word Choice subtests. These subtests are a good indicator of a child’s preparedness for learning to read and spell.

The DAS–II Working Memory cluster had moderate correlations with PAL–II Verbal Working Memory Letters and Words Composite, Sentences: Listening and Sentences: Writing Composite, and Total Verbal Working Memory Composite. The highest correlation between the DAS–II Working Memory cluster and the PAL–II was the Verbal Working Memory, Words Subtest (.60), which indicates good convergent validity. Consistent with the intercorrelations reported earlier in this section, Verbal Working Memory was moderately associated with most PAL–II subtests. This provides additional evidence of convergent validity in that the DAS–II and PAL–II Verbal Working Memory measures correlate with other measures at a similar level.

The DAS–II Processing Speed cluster was most strongly associated with RAN tasks, correlating with all the measures at the moderate level. This result provides convergent evidence for RAN measures as being related to speed of processing. Additionally, the Processing Speed Index was moderately related to Task B–Paragraph Copying, Cross-Genre Compositional and Expository Writing scores, Verbal Working Memory, and Finger Succession–Non-Dominant Hand. Children with slow processing speed may achieve lower scores on RAN, handwriting, and composing measures.

The results of the DAS–II and PAL–II RW concurrent validity study provide convergent evidence for PAL–II Working Memory and Handwriting Total Time Composite measures. Additionally, verbal skills were identified as being an important component of many PAL–II subtests from which it can be inferred that children with language delays or disorders may produce low scores on the PAL–II. DAS–II non-verbal and spatial abilities are not associated with PAL–II RW Working Memory Subtest and Handwriting Total Time Composite.
Table V.8 presents correlation coefficients, means and standard deviations for the DAS–II clusters and composites and the PAL–II M subtests. The DAS–II Verbal and Nonverbal cluster were moderately correlated with the Place Value and Part-Whole composite scores, but had low correlations with the other math measures. Linguistic and visual-perceptual processes are associated with performance on these subtests. The Spatial cluster had a moderate correlation with Part-Whole Fractions and Mixed Numbers and Spatial Working Memory written response. The latter correlation indicates convergent evidence for the spatial processing required in the Spatial Working Memory subtest. Oral Counting correlated moderately with the Nonverbal cluster. Overall, the Place Value and Part-Whole Relationships Subtests demonstrated the most association with the DAS–II verbal, visual, and spatial ability measures, probably because they are the most conceptual measures on the PAL–II M.

The DAS–II School Readiness cluster correlated moderately with Oral Counting and RAN–Single Digits. These PAL–II measures are associated with early math skills needed for adequate academic preparedness. The DAS–II Working Memory cluster was moderately associated with a number of PAL–II M subtests, indicating general working memory skills are important in math functioning. The Spatial Working Memory subtest had a moderate correlation with DAS–II Working Memory cluster, but the Quantitative Working Memory had a low correlation. Processing Speed was correlated with RAN/RAS and Numeral Writing. These are similar to the results reported for the PAL–II RW battery and the DAS–II Processing Speed cluster.

The results indicate relatively low correlations between cognitive abilities and PAL–II M measures. The Place Value and Part-Whole Relationships Subtests had the highest correlations with the DAS–II suggesting that these subtests are the most conceptual of PAL–II M measures.

**Wechsler Nonverbal Scale of Ability**

The WNV is a nonverbal measure of general ability. Subtests are designed to eliminate or minimize verbal requirements. In addition to subtest scaled scores, the WNV yields full scale scores for the four- and two-subtest batteries. The PAL–II and WNV were administered to 72 children in K–Grade 6. Detailed demographic information is provided in Table V.6. The instruments were administered with a testing interval of 85–502 days and a mean test interval of 264 days. Table V.9 presents the correlations, means, and standard deviations for the PAL–II RW subtests and the WNV subtest and full scale scores.

The WNV subtests include Matrices, Object Assembly, Coding, Recognition (i.e., Visual Working Memory) and Spatial Span (i.e., Spatial Working Memory). It is anticipated that the WNV will not have many moderate or high correlations with PAL–II RW subtests given the verbal nature of most of the PAL–II RW tasks. The results confirmed that the two tests are generally unrelated.

Table V.10 presents the correlations, means, and standard deviations for the PAL–II M subtests and the WNV subtest and full scale scores. Similar to the results obtained between the WNV and the PAL–II RW battery, there were very small correlations between the two tasks. This is unexpected in some respects given the results reported for the DAS–II and PAL–II math study. These results do provide convergent evidence for the PAL–II Spatial Working Memory Task, which moderately correlated with WNV Spatial Span. In addition, the RAN–Single Digits and –Double Digits and RAS Rate Change score related to the WNV Coding subtest, a measure of grapho-motor processing speed.

**Correlation of the PAL–II With a Measure of Neuropsychological Functioning**

The following study evaluates the degree of relationship between PAL–II and neuropsychological functioning as assessed by the NEPSY–II. The NEPSY–II has measures of Attention, Executive Functioning, Language Processing, Memory, Visual-Spatial Processing, and Sensorimotor processing. The PAL–II and NEPSY–II were administered to 55 children in K–Grade 6. Detailed demographic information is provided in Table V.6. The instruments were administered with a
testing interval of 1–316 days and a mean test interval of 114 days. Tables V.11a–V.11c present the correlations, means, and standard deviations for the PAL–II RW subtests and the NEPSY–II selected subtests by domain.

The NEPSY–II Attention and Executive domain includes measures of auditory attention, concept formation, cognitive flexibility, inhibitory control, self-monitoring, and planning. In this domain, the NEPSY–II Inhibition, Inhibition Switching, and Inhibition Total Errors had numerous moderate correlations with PAL–II measures. The moderate relationship with RAS indicates concurrent validity of RAS as a measure of cognitive flexibility. Inhibitory control is an important construct to consider in relation to many PAL–II RW measures, particularly Pseudoword Decoding, Handwriting, Narrative Compositional Fluency, and Working Memory subtests, which had moderate correlations with the NEPSY–II Inhibition subtest.

NEPSY–II Language measures had moderate correlations with multiple PAL–II RW subtests. The NEPSY–II Comprehension of Instructions subtest, a receptive language measure, had moderate correlations with PAL–II Phonological Coding, Morphological Decoding, Morphological/Syntactical Coding, and Verbal Working Memory Composites. NEPSY–II language measures also had a moderate relationship with PAL–II Sentence Sense. NEPSY–II Phonological Processing had a moderate correlation with the PAL–II Phonemes subtest, showing convergent validity between the measures. NEPSY–II Phonological Coding had a moderate correlation with PAL–II Phonological Coding, Morphological Decoding, Morphological/Syntactical Coding, and Verbal Working Memory Composites. NEPSY–II Phonological Processing was also moderately correlated with PAL–II Pseudoword Decoding, Sentence Sense, Expository Note Taking and Cross-Genre Compositional and Expository Writing Total Complete Sentences Score. The NEPSY–II Speeded Naming Task had a high correlation with PAL–II RAN Rate Change score. It also correlated with the PAL–II RAN Composite and Letters Subtest. As previously noted in the DAS–II study, linguistic skills are an important component of the PAL–II RW battery.

NEPSY–II Design Memory was not related to the PAL–II RW subtests. The Narrative Memory subtest, which is a measure of short-term verbal memory, had moderate correlations with PAL–II Phonological Coding, Morphological Decoding, and Verbal Working Memory Composites. This subtest was also related to PAL–II Pseudoword Decoding, Expository Report Writing Organizational Quality, and Cross-Genre Compositional and Expository Writing Total Complete Sentences. Verbal memory facilitates writing skills. The NEPSY–II Word Interference Repetition and Recall task is a measure of verbal working memory and freedom from interference effects in verbal working memory. The Repetition score had moderate correlations with the PAL–II Orthographic Coding, Morphological/Syntactical Coding, and Verbal Working Memory Composites. The Recall task also correlated with PAL–II Verbal Working Memory Sentences: Listening and Sentences: Writing Composite. This result indicates convergent validity for the PAL–II Verbal Working Memory Subtests.

The NEPSY–II motor did not correlate with the PAL–II RW subtests. Motor skills have long been known to be more task-specific than cognitive skills. Even within the PAL–II the motor subtests differ as to whether they are related to reading, writing, and math. There was a moderate association between NEPSY–II Block Construction and PAL–II Task B–Paragraph Copying and between NEPSY–II Working Memory and PAL–II Morphological/Syntactical Coding. This association is likely due to the speeded component to the NEPSY–II Block Construction task. Otherwise, there is little association between the NEPSY–II visual-perceptual tasks and PAL–II RW measures.

Tables V.12a–V.12c present the correlations, means, and standard deviations for the PAL–II M subtests and NEPSY–II selected subtests by domain. The NEPSY–II measures of attention and executive functioning were generally unrelated to PAL–II M subtests. The NEPSY–II Clocks subtest had a moderate correlation with the PAL–II Place Value and Part-Whole Fractions.
measures but not the Part-Whole Time task. The NEPSY–II Clocks subtest focuses more on executive functioning while PAL–II Part–Whole Time focuses on math-related time skills. PAL–II RAN–Single Digits and –Double Digits were moderately correlated with the NEPSY–II Inhibition subtest. PAL–II Finding the Bug was associated with NEPSY–II Inhibition Total Errors, which is an indicator of self-monitoring. This provides an indication of convergent validity for executive function measures.

NEPSY–II Language measures had low correlations with the PAL–II M subtests, providing discriminant validity. The exceptions were for PAL–II Part-Whole Fractions and Mixed Numbers, Place Value Verbal Response, and Quantitative Working Memory with NEPSY–II Receptive Language and Phonological Processing. NEPSY–II Speeded Naming was moderately related to PAL–II Numeral Writing and RAN tasks. NEPSY–II visual memory was not correlated with PAL–II M subtests. NEPSY–II verbal immediate memory and working memory measures had low correlations with PAL–II M subtests.

NEPSY–II motor measures had low correlation with PAL–II M measures. In the Visual-Spatial Domain, NEPSY–II Arrows (i.e., a measure of ability to determine direction of angles) was correlated with PAL–II Oral Counting and Place Value, both of which involve the dimension of direction in time or space. NEPSY–II Geometric Puzzles were also moderately correlated with PAL–II Oral Counting. Otherwise, NEPSY–II visual-spatial tasks were not associated with PAL–II M scores.

In normally developing children, NEPSY–II appears to provide broadband assessment of neurodevelopmental processes that are generally not related to the PAL–II narrowband assessment of neuropsychological and cognitive processes related to reading, writing, and math. Each assessment provides a different kind of information.

Evidence Based on Consequences of Testing

Psychological and educational tests are commonly administered to derive some benefit from the test results for the child. Tests assist professionals and institutions in meeting professional and legal standards governing diagnosis and classification, program and institutional planning, and research and evaluation activities. Examiners must be alert to intended and unintended consequences of test use. They should also acknowledge intended and unintended consequences of relying on informal judgments instead of those informed from test results. Although information about the consequences of testing may influence decisions about test use (e.g., whether or not to use a test), adverse consequences do not in themselves detract from the validity of the intended test interpretations (AERA et al., 1999).
Table P.1  Percentages of the standardization sample and U.S. population by grade and parent education level

<table>
<thead>
<tr>
<th>Grade</th>
<th>Parent Education Level</th>
<th>≤11 Years</th>
<th>12 Years</th>
<th>&gt;12 Years</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>U.S. Population</td>
<td>12.7</td>
<td>23.3</td>
<td>64.0</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>12.0</td>
<td>24.0</td>
<td>64.0</td>
</tr>
<tr>
<td>1</td>
<td>U.S. Population</td>
<td>13.9</td>
<td>26.7</td>
<td>59.4</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>14.0</td>
<td>26.0</td>
<td>60.0</td>
</tr>
<tr>
<td>2</td>
<td>U.S. Population</td>
<td>11.4</td>
<td>24.7</td>
<td>63.9</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>10.0</td>
<td>25.0</td>
<td>65.0</td>
</tr>
<tr>
<td>3</td>
<td>U.S. Population</td>
<td>13.2</td>
<td>27.0</td>
<td>59.8</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>14.0</td>
<td>26.0</td>
<td>60.0</td>
</tr>
<tr>
<td>4</td>
<td>U.S. Population</td>
<td>13.7</td>
<td>27.5</td>
<td>58.8</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>17.0</td>
<td>24.0</td>
<td>59.0</td>
</tr>
<tr>
<td>5</td>
<td>U.S. Population</td>
<td>14.8</td>
<td>25.6</td>
<td>59.6</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>15.0</td>
<td>26.0</td>
<td>59.0</td>
</tr>
<tr>
<td>6</td>
<td>U.S. Population</td>
<td>13.0</td>
<td>26.5</td>
<td>60.5</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>14.0</td>
<td>25.0</td>
<td>61.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>Race/Ethnicity</th>
<th>U.S. Population</th>
<th>Standardization Sample</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>White</td>
<td>60.4</td>
<td>62.0</td>
</tr>
<tr>
<td>K</td>
<td>African American</td>
<td>14.1</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>18.6</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>6.9</td>
<td>8.0</td>
</tr>
<tr>
<td>1</td>
<td>U.S. Population</td>
<td>57.9</td>
<td>57.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>14.6</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>20.3</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7.2</td>
<td>8.0</td>
</tr>
<tr>
<td>2</td>
<td>U.S. Population</td>
<td>61.0</td>
<td>63.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>13.8</td>
<td>14.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>17.2</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>8.0</td>
<td>6.0</td>
</tr>
<tr>
<td>3</td>
<td>U.S. Population</td>
<td>55.6</td>
<td>55.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>17.3</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>19.6</td>
<td>20.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7.5</td>
<td>8.0</td>
</tr>
<tr>
<td>4</td>
<td>U.S. Population</td>
<td>56.8</td>
<td>58.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>17.1</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>18.7</td>
<td>19.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7.4</td>
<td>7.0</td>
</tr>
<tr>
<td>5</td>
<td>U.S. Population</td>
<td>58.7</td>
<td>59.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>15.7</td>
<td>17.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>19.0</td>
<td>19.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>6.6</td>
<td>5.0</td>
</tr>
<tr>
<td>6</td>
<td>U.S. Population</td>
<td>61.0</td>
<td>61.0</td>
</tr>
<tr>
<td></td>
<td>African American</td>
<td>14.7</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>Hispanic</td>
<td>17.0</td>
<td>16.0</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>7.3</td>
<td>8.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Grade</th>
<th>Region</th>
<th>North Central</th>
<th>Northeast</th>
<th>South</th>
<th>West</th>
</tr>
</thead>
<tbody>
<tr>
<td>K</td>
<td>U.S. Population</td>
<td>21.9</td>
<td>18.8</td>
<td>35.5</td>
<td>23.8</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>21.0</td>
<td>18.0</td>
<td>35.0</td>
<td>26.0</td>
</tr>
<tr>
<td>1</td>
<td>U.S. Population</td>
<td>21.8</td>
<td>16.6</td>
<td>37.5</td>
<td>24.1</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>23.0</td>
<td>18.0</td>
<td>35.0</td>
<td>24.0</td>
</tr>
<tr>
<td>2</td>
<td>U.S. Population</td>
<td>22.3</td>
<td>18.4</td>
<td>35.3</td>
<td>24.0</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>22.0</td>
<td>20.0</td>
<td>34.0</td>
<td>24.0</td>
</tr>
<tr>
<td>3</td>
<td>U.S. Population</td>
<td>21.5</td>
<td>17.3</td>
<td>36.3</td>
<td>24.9</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>22.0</td>
<td>16.0</td>
<td>36.0</td>
<td>26.0</td>
</tr>
<tr>
<td>4</td>
<td>U.S. Population</td>
<td>21.9</td>
<td>17.4</td>
<td>35.5</td>
<td>25.2</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>22.0</td>
<td>17.0</td>
<td>36.0</td>
<td>25.0</td>
</tr>
<tr>
<td>5</td>
<td>U.S. Population</td>
<td>23.7</td>
<td>18.4</td>
<td>33.1</td>
<td>24.8</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>19.0</td>
<td>21.0</td>
<td>34.0</td>
<td>26.0</td>
</tr>
<tr>
<td>6</td>
<td>U.S. Population</td>
<td>22.1</td>
<td>18.4</td>
<td>35.3</td>
<td>24.2</td>
</tr>
<tr>
<td></td>
<td>Standardization Sample</td>
<td>22.0</td>
<td>17.0</td>
<td>38.0</td>
<td>23.0</td>
</tr>
</tbody>
</table>

### Table V.6  Demographic data for PAL–II validity studies with other measures

<table>
<thead>
<tr>
<th></th>
<th>WNV (%)</th>
<th>DAS–II (%)</th>
<th>NEPSY–II (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>N</strong></td>
<td>72</td>
<td>90</td>
<td>55</td>
</tr>
<tr>
<td><strong>Grade</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>K</td>
<td>13.89</td>
<td>15.55</td>
<td>9.09</td>
</tr>
<tr>
<td>1</td>
<td>12.50</td>
<td>11.11</td>
<td>5.45</td>
</tr>
<tr>
<td>2</td>
<td>15.28</td>
<td>17.78</td>
<td>21.82</td>
</tr>
<tr>
<td>3</td>
<td>16.66</td>
<td>16.67</td>
<td>21.82</td>
</tr>
<tr>
<td>4</td>
<td>15.28</td>
<td>12.22</td>
<td>18.18</td>
</tr>
<tr>
<td>5</td>
<td>12.50</td>
<td>7.78</td>
<td>12.73</td>
</tr>
<tr>
<td>6</td>
<td>13.89</td>
<td>18.89</td>
<td>10.91</td>
</tr>
<tr>
<td><strong>Sex</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>44.44</td>
<td>44.44</td>
<td>49.09</td>
</tr>
<tr>
<td>Male</td>
<td>55.56</td>
<td>55.56</td>
<td>50.91</td>
</tr>
<tr>
<td><strong>Race/Ethnicity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>55.56</td>
<td>51.11</td>
<td>49.09</td>
</tr>
<tr>
<td>African American</td>
<td>16.67</td>
<td>13.33</td>
<td>20.00</td>
</tr>
<tr>
<td>Hispanic</td>
<td>19.44</td>
<td>31.11</td>
<td>25.46</td>
</tr>
<tr>
<td>Other</td>
<td>8.33</td>
<td>4.45</td>
<td>5.45</td>
</tr>
<tr>
<td><strong>Parent Education</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≤11 years</td>
<td>15.28</td>
<td>23.33</td>
<td>27.27</td>
</tr>
<tr>
<td>12 years</td>
<td>27.78</td>
<td>28.89</td>
<td>21.82</td>
</tr>
<tr>
<td>&gt;12 years</td>
<td>56.94</td>
<td>47.78</td>
<td>50.91</td>
</tr>
<tr>
<td><strong>Region</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>North Central</td>
<td>26.39</td>
<td>24.44</td>
<td>14.55</td>
</tr>
<tr>
<td>Northeast</td>
<td>18.05</td>
<td>25.56</td>
<td>29.09</td>
</tr>
<tr>
<td>South</td>
<td>26.39</td>
<td>31.11</td>
<td>29.09</td>
</tr>
<tr>
<td>West</td>
<td>29.17</td>
<td>18.89</td>
<td>27.27</td>
</tr>
</tbody>
</table>

**Note.** Except for N count, data are reported as percentages. WNV = Wechsler Nonverbal Scale of Ability; DAS–II = Differential Ability Scales–Second Edition.