792378 HITOMETER

Dunmont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Test Reviews and Commentary

---

**WRAML** Comments on the WRAML factor structure

**The Process Assessment of the Learner – Test Battery for Reading and Writing (PAL-RW)** A short review with comments.

**Comprehensive Test of Phonological Processing (CTOPP)** **Description** of this new test of Phonological Awareness

For a comprehensive **review of the CTOPP by Jim Lennon and Christine Slesinski**

**KBIT vs. WISC-III** A research paper that compared the results obtained from a sample of 600 children administered both the K-BIT and the WISC-III.

**Computerized CPT Review** Reviews of the Conners', The TOVA, and the IVA Continuous Performance Test

**CPT Comparison Table**

**SENSITIVITY, SPECIFICITY, AND POSITIVE AND NEGATIVE PREDICTIVE POWER FOR CONTINUOUS PERFORMANCE TESTS** Knowing that a test is reliable and valid sometimes just isn't enough. The authors review some other issues important in the diagnosis of specific disorders.

**Software Review: A Comparison of Five Interpreter/Report Writers** Authors reviewed 5 commercial software programs that are sold as "report writers." Comparisons between the software as well as pros and cons are presented.

**FBA software** A review of two software programs (!Observe and Behavior Observation Assistant) originally published in the Communiqué.

**LET-II** A review of the Learning Efficiency Test-Second Edition

**KAIT** A review of the Kaufman Adolescent and Adult Intelligence Test

**TOMAL** A review of the Test of Memory and Learning

**ROCF** The Rey-Osterreith Complex Figure Task (Test) has been used quite a bit, yet there are some problems associated with the task. Read this interesting review and critique. Our thanks to Valerie Carter who wrote this piece.

**Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA)** A review of the Mini-Battery of Achievement

**Short Form IQ Tests** A discussion by Cisco and Eggbert about the utility of short form IQ tests.

> **How to Create a Short Form Test** Description of how to generate a deviation score for combinations of subtests.

**As the Block Turns: Interpretation of block rotations on block design subtests** Why do kids rotate those blocks anyway? Possibly some answers.

**School Psychologists or Soothsayer?** What can we really say from the result of our tests

**Norms** An interesting bit of history that highlights the need to check the normative data provided by test publishers. This involves the Halstead-Reitan test.

**Test Descriptions** A brief description of a number of commonly used assessment tools.

**How To Review A Test** An outline that may be useful in evaluating a test before using it. Since everyone knows that "If the test publisher publishes a test 'It must be good'" , John Willis and I present this outline in an effort to show how to prove (or disprove) that notion.

**Tests Measuring Aspects of Phonological Awareness** Comparisons of several new and old test of Phonological Awareness

**Comprehensive Tests of Nonverbal Intelligence (CTONI)** Short and to the point description by John Willis of this test.

**ORAL AND WRITTEN LANGUAGE SCALES (OWLS)** Review of the OWLS Written Scale by John Willis.


CLICK HERE

**Content on these pages is copyrighted  by Dumont/Willis © (2001) unless otherwise noted.**

792382
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# WRAML

A real example of how factor analysis can change the way one understands and interprets a test is the Wide Range Assessment of Memory and Learning (WRAML; Adams & Sheslow, 1990). On this test, 3 memory scales (Verbal, Visual, and Learning), each consisting of 3 subtests, are provided. Technical data reported in the administration manual for the test (pg 93) provide the results of a principal component analysis with a varimax rotation. Table XXX below summarizes that data in two ways – first with the subtests grouped according to factors proposed by the WRAML authors and second by factor loading weights.

WRAML Scale Compositions (9 & older)

**WRAML structure as proposed by authors**

| Verbal Scale | | Visual Scale | | Learning Scale | |
|---|---|---|---|---|---|
| Number/Letter | (.837) | Design Memory | (.720) | Sound Symbol | (.638) |
| Sentence Memory | (.749) | Picture Memory | (.674) | Verbal Learning | (.648) |
| Story Memory | (.196) | Finger Windows | (.584) | Visual Learning | (.401) |

**WRAML structure by factor analysis**

| Verbal Scale | | Visual Scale | | Learning Scale | |
|---|---|---|---|---|---|
| Number/Letter | (.837) | Design Memory | (.720) | Sound Symbol | (.638) |
| Sentence Memory | (.749) | Picture Memory | (.674) | Verbal Learning | (.648) |
| Finger Windows | (.585) | Visual Learning | (.583) | Story Memory | (.695) |

As can be seen, certain subtests do not seem to "load" well on the scales into which they have been placed. The Story Memory subtest loads on the Verbal scale with a weight of .196, while loading on the Learning Factor with a weight of .695. The "alternative" factor groupings seem to, statistically, "hang together" better than the actual test groupings. The WRAML authors note, in the administration manual (pg 93) that "It was decided to keep the subtests in the factors which were theorized by the authors because of the logical consistency offered in Chapters 1 and 2. Further research could change this decision." Interestingly, Phelps (1995), in her study of the WRAML factor structure, concludes that "…data indicate that the WRAML should be revised such that the subtest placement and Index scores match empirical findings."

Phelps, L. (1990). Exploratory Factor analysis of the WRAML with academically at-risk students. *Journal of Psychoeducational Assessment, 13*, 384-390.

To see another aspect of the WRAML that John and I had issue with, see the **Sexist Story Memory**.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792383
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# The Process Assessment of the Learner – Test Battery for Reading and Writing (PAL-RW)

## A quick review - By Ron Dumont and John Willis


The Process Assessment of the Learner – Test Battery for Reading and Writing (PAL-RW; Berninger, 2001) (The Psychological Corporation), uses a variety of tasks to assess a children's development of reading and writing processes. The PAL-RW was normed in 1999-2000 on 868 individuals in grades K-6 from around the U.S. According to the author, the PAL-RW can be used to:

- ***Screen*** *by identifying students at risk for reading/writing problems;*
- ***Monitor*** *by tracking progress for students in early intervention and prevention programs; and*
- ***Diagnose*** *by evaluating the nature of reading/writing–related processing problems.*

**Complete Kit in a Box**
Includes Examiner's Manual, Stimulus Booklets, 25 Record Forms, 25 Response Forms, Stylus-Wood, Word Card, Audiotape, and Shield.
**$250.00**

The PAL-RW appears to be a good attempt at measuring the emerging skills needed for the complicated tasks of reading and writing. As a diagnostic tool for early grade school children, it appears to be quite useful. Its use with older children may be hampered by the limited number of items on certain subtests. The scores obtained by older children may accurately reflect the problems they may have in the specific area, but the lack of sufficient numbers of items limits any diagnostic or interpretive statements that can be made. The PAL-RW would be more useful with clarification of the scoring rules, as noted below.

The use of the PAL-RW to “monitor student’s progress during and after intervention” (a stated use of the test) seems problematic given the poor test-retest statistics provided in the manual.

---

***Test description:***

*The PAL-RW includes the following subtests:*

- ***Alphabet Writing*** *(speed of writing lower-case letters of the alphabet from memory in 15 seconds)*

  - ***Receptive Coding***

  *Task A [shown a word (AT) for 1 second. Then shown IT. Are the words the same?]*
  *Task B [shown a word (BAT) for 1 second. Then shown C. Is the letter in the word?]*
  *Task C [shown a word (ATE) for 1 second. Then shown ET. Are the 2 letters in the word in the correct order?]*
  *Task D [shown a word (MOTHER) for 1 second. Then shown L. Is the letter in the word?]*
  *Task E [shown a word (SOCIETY) for 1 second. Then shown EI. Are the 2 letters in the word in the correct order?]*

- ***Expressive Coding***

  *Task A [shown a word (QAST) for 1 second. Then write the word.]*
  *Task B [shown a word (LADFUST) for 1 second. Then write the third letter.]*
  *Task C [shown a word (POGDUS) for 1 second. Then write the last 3 letters.]*

- ***Rapid Automatic Naming (RAN)***

  *Rapid Letter Naming (name these letters as fast as you can)*
  *Item 1: m t g k b h r a n  Item 2: fi ps er ou*
  *Rapid Word Naming (name these words as fast as you can)*
  *dog eat of sit over*
  *Rapid Digit Naming (name these numbers as fast as you can)*
  *Item 1: 3 7 8 1 9 6 2)  Item 2: 67 89 45 73*
  *Rapid Word and Digit Naming (name these Words and Digits as fast as you can)*
  *tea eat 56 of 89 over*

- ***Note-Taking Task - A*** *(Listen to a story and take notes as it is read)*

- ***Rhyming***
  *Task A (Listen to 3 words and tell which one does not have the same sound)*
  *ball call help*

*Task B (The word is PIG. Tell me all the real words you can that rhyme with PIG.)*

- ***Syllables*** *(Hear a word (both real and made-up), say the word, now say it with a sound left out)*
  *PUTTING Say PUTTING Now say it without the PUT*
- ***Phonemes*** *(Hear a word (both real and made up), say the word, now say it with a sound left out – what sound was left out)*
  *SIT Say SIT Now say IT What sound is missing?*
- ***Rimes*** *(Say a word (real or made up)with a sound left out)*
  *Say BIKE without /b/*
- ***Word Choice*** *(Shown 3 words, indicate the one which is spelled correctly)*
  *PIG PAG PIZE*
- ***Pseudoword Decoding*** *(Read some words that are not real words)*
  *DRIY HAFFE STROC*
- ***Story Retell*** *(After being read a short story, answer questions, then retell story in own words)*
- ***Finger Sense***

  *Repetition (1 & 2) [Touch thumb to index finger 20 times (Right and left hands) scored for completion time]*
  *Succession (1 & 2) [Touch thumb to each finger 5 complete times (Right and left hands) scored for completion time]*
  *Localization (After having one finger touched out of sight, tell which finger was touched)*
  *Recognition (Each finger is assigned a number. After having one finger touched out of sight, tell what number of the finger was touched)*
  *Fingertip Writing (After having a letter "written" onto a fingertip, tell which letter was written)*
- ***Sentence Sense*** *(Read 3 sentences and tell which one makes sense)*
  *I ATE THE CAKE*
  *I EIGHT THE CAKE*
  *I ATE THE CAPE*
- ***Copying*** *(Here is a sentence (paragraph). Copy it as fast as you can)*

  *Task A THE LAZY BOY JUMPED OVER A BALL*
  *Task B (A paragraph)*
- ***Note-Taking Task B*** *(Take the notes created earlier [Note-Taking Task A] and write a paragraph based on the notes)*

---

**General Comments:**

**All scores are based upon the grade of the child tested, not the chronological age. No explanation is given for why this is so. Are these "neurodevelopmental processes" age- or grade-dependent?**

**Normative sampling seems adequate [>100 at each grade (range 105 – grade 6 to 142 – grade 1)]. Appropriate percentage comparisons to the U.S. population are evident for sex; race/ethnicity; parental education; and geographic region.**

**All scores are reported as DECILE scores. These describe which tenth of the distribution the child's performance lies in. A child's Decile score of 20 means that 20% of the general population was at or below the child's performance. The PAL divides the Decile scores into descriptive categories (10-20 - Deficient, 30-40 - At Risk, 50 - Emerging Adequate, 60-80 – Adequate, and 90-100 – Proficient).**

**Test-retest comparisons based on 86 children in Grades 1, 3, and 5 tested a second time 14 to 49 days later, show reliabilities that ranged from .61 to .92. Five measures had reliabilities below .70. Seven of the 14 tests had lower scores on retest!**

**Criterion-related validity studies with individually administered tests varied greatly in sample size (WIAT-II, n = 120, PPVT-III, n = 19-43, VMI, n = 7-12, and CELF-III, n = 14). Despite the relatively small sample sizes in some of the validity studies, the PAL-RW generally did show expected correlations with other reading, decoding, and language tests.**

**Comparison between clinical and nonclinical samples suffer from limitations in sample size. For example, of 18 measures assessed and compared, the sampling range from an n of 3 to an n of 23. Result found significant differences ($p < .05$) for 7 of the 18 subtests**

**Examiners may wish to add some tabs to the easels since they contain multiple subtests and differing starting points.**

**The pages on the easel are sturdy, but after repeated use, they begin to tear away from the ring binder. This is especially true on the pages for which the examiner is required to flip after only 1 second of exposure. Examiners may wish to apply ring reinforcers or to apply heavy tape and repunch the holes.**

---

**Subtest comments:**

**Alphabet Writing –**

- **This test is scored for the number of correct, unique letters the child has reproduced in 15 seconds.**
- **For those children who are very young or very slow, the record form provides space to record the number of letters completed in 5 minutes. No norms are provided for this condition.**
- **Norms for Grades K (both Fall and Spring) are based upon WIAT-II standardization sampling**
- **Scoring examples are given in the manual although no explanations of how to determine "Too closed" or "Too open" are given.**
- **You do not count as correct "letters that are out of order." This seems a bit confusing. If a child writes "a b d c e f h g", how does one score it? Since the 3rd letter is incorrect, are all others out of order resulting in a score of 2? The manual does not elaborate.**

***Rapid Automatic Naming (RAN)***

- **The score for these tasks is based upon the amount of time it takes to name the letters or numbers presented. However, although the examiner keeps track of errors, the Decile score is simply based on the speed, not the accuracy of completion. Norms for errors suggest that any error at any age places the child in the Deficient or At-Risk category.**

*Note-Taking Task - A*

- **No tape is provided to the examiner leaving a wide variation of how the story can be read. The manual notes that the examiner should read in a "normal, conversational tone" similar to a class lecture. There is no emphasis on any words or parts of the passage.**
- **Scoring is done by comparing the notes taken with criteria relating to Main Ideas and Supporting Ideas. Two of the main ideas have credit for the same supporting detail. It is unclear if the child must say the supporting idea twice to get the credit or if by mentioning it once, he or she gets the credit twice.**
- **Scoring also includes scores for "Attributes" rated as Never, Rarely, Sometimes, Often, and Always Present. These attributes, and their scoring, are seemingly ill-defined, with no examples given to explain what is meant by or how to judge some of the attributes (e.g., "Notes are legible"). In the attribute section, up to 4 points are given for "Notes are accurate" (undefined), despite the fact that the Main Ideas and Supporting Ideas scoring section presumably was measuring accuracy.**

*Syllables*

- **Grades 1-3 start with item Sample 3, and number 11. If, after taking 10 items, the child has failed any two items, examiners go back and administer items 1-10. However, children in grades 4-6 start at Sample 4, and are administered only 6 items. Regardless of the number of errors, examiners do not administer earlier items? That seems like a small number of items to create a stable score.**

*Phonemes and Rimes*

- **Each subtest provides only 6 items for children in grades 4-6**

*Finger sense*

- **Succession (1 & 2) [Touch thumb to each finger 5 complete times (Right and left hands)]. Examiners are to record the "Finger order of Incorrect Sequences." This seems to be a fairly difficult task and one, like many of the supplemental recordings made by the examiner, which has no score nor any interpretive suggestions.**

*Copying [Here is a sentence (paragraph). Copy it as fast as you can.]*

- **Despite the fact that the paragraph for Task B contains both capital letters and punctuation marks, only the number of correctly copied letters is counted.**
- **The directions for scoring include "Do not count letters that are written in capitals (uppercase). Yet 14 words in the paragraph are capitalized!?**
- **No directions for how to score run-on words is given. Must the child copy the sentence and paragraph with correct spacing between the words?**

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792384
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Comprehensive Test of Phonological Processing (CTOPP)


Wagner, Torgesen, & Rashotte (Pro-Ed, 1999). The CTOPP, with different forms for ages 7 – 21and 5 – 6, uses a variety of tasks to assess a student's ability to perceive and manipulate the sounds that make up words. The CTOPP was normed in 1997-98 on 1,656 individuals around the U.S. Many of the tasks are presented by tape recordings. It includes the following subtests and scales (no rhyming, sadly):

- elision (say "blend" without saying /l/)
- blending words (what word do these sounds make? /k/ /a/ /t/)
- memory for digits (say these numbers: three, five, one, nine, six, two)
- rapid digit naming (name these numbers as fast as you can: 3 7 8 1 9 6 2)
- nonword repetition (say "dooloowheep")
- rapid letter naming (name these letters as fast as you can: m t g k b h r a n)
- rapid color naming (name these colors as fast as you)
- phoneme reversal (say "foob"; now say "foob" backwards)
- rapid object naming (name these objects as fast as you can: ??? ? ??)
- blending nonwords (what made-up word do these sounds make? /r/ /ç/ /b/)
- segmenting words (say "pit" one sound at a time)
- segmenting nonwords (say "kloop" one sound at a time)
- Phonological Awareness: elision and blending words
- Phonological Memory: memory digits and nonword repetition
- Rapid Naming: rapid digit naming and rapid letter naming
- Alternate Phonological Awareness: blending and segmenting nonwords
- Alternate Rapid Naming: rapid color naming and rapid object naming

To read a review of the CTOPP press here

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792385 HITOMETER

Dunmont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Comprehensive Test of Phonological Processing (CTOPP): Cognitive-linguistic assessment of severe reading problems

**James E. Lennon** press name to email author

New Jersey City University

Christine Slesinski

Monroe-Woodbury Central School District

Wagner, Torgesen, & Rashotte (1999) recently offered the Comprehensive Test of Phonological Processing (CTOPP) as a measure of phonological coding. This measure may be of value to school psychologists who are interested in the etiology of severe reading disorders. Rather than seeking to identify intelligence-achievement discrepancies of limited utility, assessment approaches, such as the CTOPP, that measure phonological coding abilities may help school psychologists to more accurately differentiate students with learning disabilities from students who may be experiencing academic failure as a result of other causes.

Phonological coding consists of the analysis and synthesis of phonemes (the smallest unit of recognized sounds). Beginning readers who have deficits in phonological coding seem to have difficulty naming letters of the alphabet, identifying sounds for alphabet letters, segmenting words into phonemes and syllables, and applying knowledge of letter-sound correspondence to decode words (Vellutino, et al., 1996). Phonological coding is an oral language skill. It involves analysis such as recognizing that the first sound of the word ball (/b/), can be replaced with /t/ to produce the word tall. Phonological coding abilities associated with this process of changing ball to tall include letter-sound correspondence, phonemic awareness and segmentation, and working with information in phonological memory. It also involves the synthesis of sounds into words. Since the most common forms of severe reading problems are caused by deficits in one or more aspects of phonological coding, school psychologists should consider including measures specifically designed to address this cognitive-linguistic process in their assessment of cognitive functioning.

The recently published Comprehensive Test of Phonological Processing (CTOPP; Wagner, Torgensen & Rashotte,1999) has been designed as an extension and improvement over commercially available tests of phonological coding, including the Test of Phonological Awareness (TOPA; Torgensen and Bryant, 1994), the Lindamood Auditory Conceptualization Test (LAC; Lindamood & Lindamood, 1979), and the Phonological Awareness Test (PAT, Robertson & Salter, 1995). The CTOPP provides greater extension across age ranges, has stronger psychometric properties than previous measures (Torgesen & Wagner, 1998), and will be reviewed in the paragraphs that follow.

Wagner, Torgensen & Rashotte (1999) developed the CTOPP in a manner consistent with their theoretical assumptions about the nature of phonological coding deficits. They present a three-part model, based on earlier studies in this area (e.g., Torgesen & Wagner, 1998; Wagner & Torgesen, 1987), consisting of the following:

- (a.) Phonological awareness: analysis and synthesis of the sound structure of oral language. The order of progression of phonological awareness starts with syllables and moves toward smaller units of speech sounds (Adams, 1990). Phonological awareness provides individuals with the ability to break words into syllables and component phonemes, to synthesize words from discrete sounds, and to learn about the distinctive features of words (Torgesen & Wagner, 1998).
- (b.) Phonological memory: coding information phonologically for temporary storage in working or short-term memory. Phonological short-term memory involves storing distinct phonological features for short periods of time to be "read off" in the process of applying the alphabetic principle to word identification.
- (c.) Rapid naming: efficient retrieval of a series of names of objects, colors, digits, or letters from long-term memory. Rapid naming of verbal material is a measure of the fluid access to verbal names, in isolation or as part of a series, and related efficiency in activating name codes from memory (Wagner, Torgesen, & Rashotte, 1999).

Grounded in a theory of phonological processing, supported by both empirical studies and confirmatory factor analytic findings, the CTOPP was designed to measure a student's ability in these three domains (Wagner, et al., 1997).

The CTOPP is intended to provide a reliable, valid, and standardized measure of phonological coding. The authors developed two versions of the measure, one for kindergarteners and first graders (ages 5 and 6) and the second for second graders through college students (ages 7 through 24). A total of 12 subtests (6 core and 6 supplemental) are provided. Subtests typically consist of 18 to 20 items, providing adequate floors and ceilings. The CTOPP is individually administered, and requires about 30 minutes of testing time to administer the core subtests.

The test produces three core composite scores: (a.) phonological awareness, comprised of Elision, Blending Words and Sound Matching for 5 and 6 year-olds and Elision and Blending Words for persons 7 to 24-years-old; (b.) phonological memory, consisting of Memory for Digits and Nonword Repetition for all individuals and (c.) rapid naming comprised of Rapid Color Naming and Rapid Object Naming, and Rapid Digit Naming and Rapid Letter Naming, for younger and older students respectively. Rather than relying on a single measure, the CTOPP provides two measures for each composite score, increasing the reliability of the measurement of the construct. Additional alternate measures of phonological awareness and rapid naming are provided for clinical and research interest. Interpretation involves the conversion of raw scores into percentile ranks and standard scores (mean = 10, standard deviation = 3 for subtests; mean = 100, standard deviation = 15 for composite scores). The authors offer, but caution against using, age and grade equivalent scores

The measure was normed on a stratified sample of 1,656 individuals, reflecting the demographic status of the US population in 1997. Between 76 to 155 students were included in each age range, with the greater representation in the youngest age ranges. Normative information in provided by the half-year for 5 and 6-year-olds, by the year for 7 through 17-year-olds, and for an 18 to 24-year-old composite group.

CTOPP subtests were derived from experimental tasks used in the research literature to assess phonological processing. Pilot studies allowed for extensive item and subtest analyses, including classical item analyses, item-response theory, and confirmatory factor analyses. There were careful efforts to design empirical tasks that were representative of the constructs in question.

For example, the phonological memory composite includes Memory for Digits and Nonword repetition. Though similar to the Digit Span test of the WISC-III, Memory for Digits is presented via audiocassette recorder at a faster rate of presentation (two digits per second) and with a specification of forward-only recall. These modifications were meant to stress the efficiency of the phonological loop, i.e., "brief verbatim

storage of auditory information" (Wagner, Torgesen & Rashotte, 1999, p.5), avoiding the involvement of other cognitive processes, such as rehearsal and elaboration. On Digit Span, many students "think through" the backward recall of numbers or refresh the phonological loop through rehearsal, calling on other cognitive strategies. The modifications on the CTOPP attempt a purer measure of the underlying phonological process that is not confounded by other cognitive operations.

Nonword repetition has also been shown to be a good measure of phonological memory in experimental tasks (Gathercole & Baddeley, 1990). The authors created the orthographically legitimate, or plausible English language, items such as "nirp" by randomly combining phonemes to fill slots in syllables, discarding non-pronounceable ones. This was done to avoid the possible confound of using analogies to real words, once again avoiding the use of cognitive-linguistic processes other than phonological memory (R. Wagner, personal communication, May 15, 2000). Similar to most subtests, Nonword Repetition requires the use of an audiocassette recorder to ensure standardized administration, particularly as the items become more difficult.

Measures of both analysis (Elision) and synthesis (Blending Words) are included in the composite for phonological awareness, consistent with recent factor analytic findings (Flanagan, McGrew, & Ortiz, 2000; Wagner, et al., 1997). Alternate measures of the analysis and synthesis components of phonological awareness using nonwords are offered for experimental or clinical interest. Additionally, experimental measures with unlimited ceilings are included. For example, Phoneme Reversal requires the repetition of a nonword, reversing the order of sounds, and pronouncing the resultant word. ("Say teef. Now say teef backward. " Answer: feet.) Phoneme reversal requires the coordination of phonologic, strategic, and memory processes. As noted in the manual, the measures of the CTOPP can be quite challenging for both the examiner and examinee.

Some researchers describe Rapid Naming as part of a "double deficit hypothesis" (Wolf & Bower, 1999), representing a separate category of severe reading deficits along with phonological coding deficits. However, on the CTOPP as a result of confirmatory factor analytic findings, the rapid naming composite is thought to be a component process of the phonological coding construct, correlated to other components, but containing unique variance as well (Wagner, Torgesen, & Rashotte, 1999).

The CTOPP appears to have sound technical features. Reliability estimates of internal consistency of the items are provided. The age interval alpha coefficients of the CTOPP subtests reach .80 reliability, 76% of the time, while the CTOPP composite scores reach the .80 reliability criterion. Standard errors of measurement for the composite scores are relatively low, suggesting the composite scores are reliable measures of student performance.

Reliability over time was estimated by the test-retest method, and ranged from .70 to .97 for individual subtests and .78 to .95 for composite scores. Measurement reliability is improved by using more than a single subtest to report composite scores.

Validity information is offered in the form of (a.) a detailed discussion of the rationale used in selecting items and subtest format, (b.) conventional item analysis and response theory modeling and (c.) logistic regression and delta scores to detect bias. Little or no bias in the groups investigated was reported. Item discrimination and item difficulty statistics reach acceptable levels. Criterion-related validity is reported between concurrent measures, such as the Lindamood Auditory Conception Test, and predictive measures, such as the Woodcock Reading Mastery Test – R (Word Attack and Word Identification subtests). Finally, construct validity is reported in the form of confirmatory factor analysis and studies of age group differentiation.

**Using the CTOPP: Domain-specific deficits vs. IQ/Achievement discrepancies**

Now consider the case of an entering student, Hannah, a five-year, eight-month old kindergarten student in the same school district. Hannah, the oldest of three children, currently lives at home with both parents and her younger siblings. Hannah was born following a full-term, uncomplicated pregnancy and achieved developmental milestones within normal limits. Between the ages of one and three, Hannah suffered from recurrent ear infections. At the age of three, she underwent surgery to remove her adenoids and have drainage tubes placed in her ears. Prior to attending kindergarten, Hannah attended an academically oriented preschool for two years. Based on her below average performance on a kindergarten screening measure, Hannah was placed in a regular kindergarten classroom, but referred for remedial reading instruction.

After the first marking period, Hannah's classroom and remedial teachers referred her to the CSE. At the time of the referral, her teachers described Hannah as an intelligent student who had made a good social transition to kindergarten. They noted that she was friendly, eager to please, and attentive in class. Despite this, they expressed concerns that Hannah just didn't seem to be "getting it" when it came to reading. They noted that she could identify only 8 letters consistently, did not evidence knowledge of letter-sound correspondence, had difficulty rhyming words, and was unable to identify sounds in spoken words. While she could write her name, they described this ability as a "rote-learned" skill. Psychological testing revealed average intelligence (WPPSI Full Scale IQ = 105), with no significant inter-subtest scatter. Commensurate with her intelligence, Hannah's achievement was measured to be in the average range (WJ-R Broad Reading standard score = 95; WJ-R Broad Written Language standard score = 101.) Despite average intelligence, average achievement, and seemingly appropriate early educational experiences, Hannah was having considerable academic difficulty in her kindergarten classroom. Once again the school psychologist was left to wonder what was going on. The absence of an aptitude-achievement discrepancy made it difficult to attribute Hannah's reading problems to a learning disability. Rather than waiting to see if a learning disability "developed," the school psychologist wanted to take a closer direct look at the cognitive-linguistic operations that underlie beginning reading.

In this case, the teachers suspected that the student's reading difficulties resulted from learning disabilities, prompting referrals to the CSE. In turn, the school psychologist attempted to diagnose learning disabilities in the student by identifying an aptitude-achievement discrepancy. However, the search for aptitude-achievement discrepancies left important questions unanswered. While many IQ tests do not address the processes that are associated with significant reading difficulties (Fletcher, et al, 1998; Flanagan, McGrew, & Ortiz, 2000), one might argue that IQ tests are necessary to rule out basic process disorders. However, Stanovich has argued that the concept of unexpected intelligence-achievement discrepancies has "led us astray'" (1991,p.7). He suggests cognitive-linguistic deficits are not necessarily restricted to students in the average to above average range and argues that the measurement of such deficits must be a domain-specific process.

School psychologists search for sources of severe reading problems in various ways. Typically, the search involves identifying students who have significant aptitude-achievement discrepancies as learning disabled. However, serious concerns have been raised about the validity and reliability of this practice (for cogent summaries, see Fletcher, Francis, Shaywitz, Lyon & Shaywitz, 1998; Siegel, 1989; Stanovich, 1991; Vellutino, Scanlon, & Lyon, 2000). These studies, in part, question the relevance of administering global measures of intelligence, which do not tap reading related cognitive abilities, to students suspected of having learning disabilities (see also Flanagan, McGrew, & Ortiz, 2000 for a related discussion). Converging research evidence strongly suggests that the most common forms of severe reading problems are caused by deficits in one or more aspects of phonological coding, a cognitive linguistic ability (Morris, et al., 1998; Torgesen & Wagner, 1998; Vellutino, et al., 1996). Deficits in phonological coding distinguish between average and deficient beginning readers, and predict which deficient readers will demonstrate a limited response to instruction (Vellutino, et al., 1996).

How then might school psychologists best gain diagnostic information about the cognitive processes underlying severe reading problems? How can Hannah's problems be more clearly understood? Many educational researchers are suggesting that the direct assessment of phonological coding is an appropriate avenue of inquiry, because deficits in this area seem to serve as a bottleneck, impeding the development of robust reading skills (Morris, et al., 1998).

Returning to our case studies, Hannah received the core version of the CTOPP appropriate for their chronological age. Her composite scores were all below average (Phonological Awareness, 75; Phonological Memory, 80; Rapid Naming, 79). Her scores were consistently below average on all of the core subtests. Since the opportunity for instruction should precede disability determinations (Vellutino, et al.,1996), Hannah received phonological segmentation training within a balanced, intensive remedial program (see Lennon & Slesinski, 1999; Wagner, et al., 1998; Vellutino, et al., 1996, for discussions of phonological processing and reading).

Hannah did not make as much relative improvement after 10 weeks as was hoped for. She could identify more alphabet letters than when first seen, but continued to have difficulty associating the appropriate letter sounds. In oral language skills she had difficulty segmenting compound words into syllables and syllables into phonemes. After 20 weeks of remediation, the Committee on Special Education concluded that Hannah was a difficult-to-remediate child, who would benefit from being followed in a formal manner by the Committee. Additional medical information about transient, recurrent ear infections was sought. She was identified as having a learning disability, but was continued in a similar remedial program because progress was noted in both phonological awareness and word reading after 20 weeks.

## Conclusion

School psychologists may find the CTOPP challenging to administer. Familiarity and practice in using the audiocassette recorder is needed, since many of the subtests required its use and because many items are novel constructions. School psychologists may not be familiar with listening for subtle distinctions in phoneme repetition, for example. While experienced examiners may be tempted to discard the audiocassette recorder, standardization requires its use. Half-year age norms were helpful in Hannah's case, but only year-level norms are available for 7-year-olds and older. The face sheet and protocol are laid out logically, but do not provide enough space to record the examinee's errors for subsequent clinical interest. The manual provides a discussion of severe discrepancies and guidance in computing difference scores, but little direction as to the meaning of such discrepancies. This section would probably be best omitted, particularly in light of the need for a more complete discussion of discrepancy scores as noted above.

Overall the CTOPP appears to be an example of a theory-based, well-researched instrument of a domain-specific aspect of cognitive-linguistic functioning. It provided important information regarding processes that underlie beginning reading skills, and when used in conjunction with curriculum-based measures, trial teaching, and other formal assessment information is likely to aid in understanding the problems some children have learning to read.

## References

Adams, M. J. (1990). Beginning to read: Thinking and learning about print. Cambridge, MA: MIT Press.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). The Wechsler . intelligence scales and Gf-Gc theory: A contemporary approach to interpretation. Boston: Allyn and Bacon.

Fletcher, J. M., Francis, D. J., Shaywitz, S. E., Lyon, G. R., Foorman, B. R., Steubing, K. K., & Shaywiz, B. A. (1998). Intelligent testing and the discrepancy model for children with learning disabilities. Learning Disabilities Research and Practice, 13, 186-203.

Gathercole, S. E., & Baddeley, A. D. (1990). Phonological memory deficits in language disordered children: Is there a causal connection? Journal of Memory and Language, 29, 336-360.

Lennon, J. E., & Slesinski, C. (1999). Early intervention in reading: Results of a screening and intervention program for Kindergarten students. School Psychology Review, 28, 353-364.

Lindamood, C., & Lindamood, P. (1971). Lindamood auditory conception test. Austin, TX: PRO-Ed.

Morris, R. D., Stuebing, K. K., Fletcher, J. M., Shaywitz, S. E., Lyon, G. R., Shankweiler, D. P., Katz, L., Francis, D. J., & Shaywitz, B. A. (1998). Subtypes of reading disability: Variations around a phonological core. Journal of Educational Psychology, 90, 347-373.

Robetson, C., & Salter, W. (1995). Phonological Awareness Test. East Moline, IL: LinguiSystems.

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. Journal of Learning Disabilities, 22, 469-479.

Stanovich, K.E. (1991). Discrepancy definitions of reading disability: Has intelligence led us astray? Reading Research Quarterly, 26, 7-27.

Torgesen, J. K., & Bryant, B. R. (1994). Test of phonological awareness. Austin, TX: PRO-ED.

Torgesen, J. K., & Wagner, R. K. (1998). Alternative diagnostic approaches for specific developmental reading disabilities. Learning Disabilities Research and Practice, 13, 220-232.

Vellutino, F. R., Scanlon, D. M., & Lyon, G. R. (2000). Differentiating between difficult-to-remediate and readily remediated poor readers: More evidence against the IQ-Achievement discrepancy definition of reading disability. Journal of Learning Disabilities, 33, 223-238.

Vellutino, F. R., Scanlon, D. M., Sipay, E. R., Pratt, A., Chen, R., & Denckla, M. B. (1996). Cognitive profiles of difficult-to-remediate and readily remediated poor readers: Early intervention as a vehicle for distinguishing between cognitive and experiential deficits as basic causes of specific reading disability. Journal of Educational Psychology, 86, 601-638.

Wagner, R. K., & Torgesen, J. K. (1987). The nature of phonological processing and its causal role in the acquisition of reading skills. Psychological Bulletin, 101, 192-212.

Wagner, R. K., Torgesen, J. K., & Rashotte, C. A. (1999). Comprehensive test of phonological processing. Austin, TX: PRO-ED.

Wagner, R. K., Torgesen, J. K., Rashotte, C. A., Hecht, S. A., Barker, T. A., Burgess, S. R., Donahue, J., & Garon, T. (1997). Changing

relations between phonological processing abilities and word-level reading as children develop from beginning to skilled readers: A 5-year longitudinal study. Developmental Psychology, 33, 468-479.

Wolf, M., & Bowers, P. G. (1999). The double-deficit hypothesis for developmental dyslexias. Journal of Educational Psychology, 91, 415-438.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792386 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# HOW WELL DOES THE K-BIT PREDICT WISC-III RESULTS?

## C. L. Boyd and Ron Dumont

**A similar article appeared in the NASP Communiqué, 24, 6, 24**

The Kaufman Brief Intelligence Test (K-BIT) is recommended by its publisher for use as a general intellectual screening measure. For this reason, school personnel may consider using the K-BIT as part of screening procedures prior to referring a student for an evaluation by a school psychologist. Considering the popularity of the Wechsler Intelligence Scale for Children-Third Edition (WISC-III), many school psychologists may choose to use the WISC-III as a part of a student's comprehensive evaluation. School psychologists may very well be interested in knowing how accurately results from the K-BIT predict results on the WISC-III.

Since the K-BIT was published prior to the publication of the WISC-III, the K-BIT manual is only able to present correlation studies comparing the results of the K-BIT and the WISC-III's predecessor, the Wechsler Intelligence Scale for Children-Revised (WISC-R). To establish construct validity for the K-BIT, the WISC-R IQ scores for 35 normal children, ages 6-15 were compared to the K-BIT standard scores. Results of that study suggested adequate correlations between the Vocabulary, Matrices, and Composite IQ's and the Verbal, Performance, and Full Scale IQ's respectively. Correlations for the Vocabulary vs. Verbal IQ ranged from .54 to .78 while the correlations for the Matrices vs. Performance IQ ranged from .48 to .56. Correlations for the K-BIT Composite vs. the WISC-R Full Scale IQ ranged from .63 to .80. While constrained by small sample sizes, those studies revealed that the K-BIT yields a lower IQ Composite score than the WISC-R Full Scale IQ but that correlations of substantial magnitude exist between WISC-R and K-BIT results. This finding suggests that using results from the K-BIT to predict WISC-R results might be feasible. However, is it equally feasible to assume that results of the K-BIT will predict WISC-III results? Several letters to the editor appeared in Communiqué within the past year reporting considerable differences observed between prior WISC-R results and WISC-III results. Those observations seemed inconsistent with studies reported in Chapter 5 of the WISC-III manual comparing results from the WISC-R and the WISC-III. However, Dr. Lawrence Weiss (1995) reviewed 22 studies comparing WISC-R to WISC-III scores. He found "Taken together, these findings are consistent with the expected rate of change in intelligence scores over time. The expected rate of change is approximately 1/3 point per year (Flynn, 1987). Because the WISC-R and WISC-III were normed 17 years apart, there should be approximately -5.57 points difference between the respective FSIQs. The obtained difference of -5.69 points across 22 studies is remarkably similar to expectation."

To determine the relationship between results from the K-BIT and the WISC-III, protocols from 613 students referred to the Psychological Services Department of The School Board of Polk County, Florida from August 1993 through January 1994 were examined. The Polk County public schools serve a student population of about 78,000 in a geographical area approximately the size of the state of Rhode Island. While traditionally a rural county, Polk County is similar to other areas of Florida in experiencing rapid population growth and urbanization, being influenced by the adjacent metropolitan Orlando and Tampa Bay areas as well as by nearby Disney World. Interestingly, the student population in Polk County is remarkably close to national demographics in many important characteristics, including racial and ethnic

composition. Most of the students included in this study were referred for school psychological evaluations to aid in determining possible eligibility for special education programs (in descending order of predominance, Specific Learning Disabilities, Gifted, Mentally Handicapped, and Emotionally Handicapped). Some of the students were referred for an educational evaluation at parent request to determine learning strengths and weaknesses without suspicion of potential special education eligibility. None of the referred students were currently identified as eligible for any special education program, except that some students may have been receiving consultative or part-time (no more than 1.5 hours per week) services from the Speech and Language program for articulation or language therapy. Some of the referred students were receiving services from a federally-funded Chapter 1 reading reinforcement program.

The K-BIT was administered to all students in this study before referral for an individual school psychological evaluation. K-BIT testing was usually done by a school guidance counselor or other specialist with experience in administering screening assessments. Before using the K-BIT for this purpose, school personnel were trained through a series of inservice programs on the administration and scoring of the K-BIT. These inservice programs were provided to assure a minimum level of competence in use of the K-BIT. The instructors for these inservice programs included an expert provided by the publisher of the K-BIT and a Polk County school psychologist with considerable experience in using the K-BIT.

The WISC-III was administered in subsequent evaluations by Polk County school psychologists. Each of the 26 school psychologists contributing cases to this study are certified in Florida as a Specialist in School Psychology and have considerable experience in administering and scoring the Wechsler scales. Each psychologist completed a mandatory inservice training program on the administration, scoring, and interpretation of the WISC-III, including both an educational component (a three-hour workshop) and a supervised practicum (observation of a WISC-III administration, a critique of the administration, and a review of the scoring of the case).

From the available sample of 613 students, two sets of data were generated. For 200 students, data was provided for each of the 3 global scores available from the tests (WISC-III Verbal, Performance, and Full Scale; K-BIT Vocabulary, Matrices, and Composite), while for the remaining 413 students, only Full Scale and Composite scores were available.

Correlation coefficients were calculated for the 200-student subset where additional scores were available (K-BIT Vocabulary and Matrices plus WISC-III Verbal IQ and Performance IQ scores).

**Means, Standard deviations, and Correlation between WISC-III and K-BIT scales**

| | | | K-BIT Scale | | |
|---|---|---|---|---|---|
| | Mean | SD | Vocabulary | Matrices | Composite |
| WISC-III | | | | | |
| Verbal IQ | 95.45 | 19.51 | .82 | .61 | .77 |

| | | | | | |
|---|---|---|---|---|---|
| Performance IQ | 94.94 | 20.44 | .64 | .68 | .68 |
| Full Scale IQ | 94.68 | 20.44 | .77 | .75 | .83 |
| | | K-BIT mean | 96.8 | 100.4 | 98.6 |
| | | SD | 20.4 | 19.2 | 19.4 |

The K-BIT Vocabulary correlated better with the Wechsler Verbal IQ than with the Performance IQ (mean .82 versus .64), while the Matrices subtest correlated about equally with the Wechsler Performance IQ than the Wechsler Verbal IQ (mean .61 versus .68). The WISC-III Full Scale IQ correlated highly (.83) with the K-BIT Composite score and had a mean difference of approximately 4 points, with the K-BIT producing the higher score. Pearson product-moment correlations for the K-BIT IQ Composite scores and the WISC-III Full Scale IQ scores were additionally calculated for each of the 613 children separated by age group. Inspection of those correlations show them all to be acceptably high, ranging from .73 to .88 (average .83), suggesting that a strong, positive relationship exists between the results from the K-BIT and the WISC-III at each of the separate age levels.

Although this study found high correlations between the K-BIT and the WISC-III, one of the purposes for administering the K-BIT was to predict the students' current level of cognitive functioning that would be determined by the classification obtained when the entire WISC-III was administered. The mean K-BIT composite (98.6) was found to be approximately 4 points higher than the mean WISC-III FSIQ (94.2). This difference is significant ($t[612] = 9.66$, $p =<.0001$). Each student's intelligence classification obtained from the WISC-III was compared to the classification obtained from the K-BIT. Of the 613 children re-tested, 263 classification labels (43%) were found to be unchanged using the K-BIT. The K-BIT had, however, underestimated the classification of 99 children (16%) while overestimating the classification of 251 others (41%). Kaufman (1990) notes, "Probably clinicians and researchers place too much weight on the 'misclassification index', because so called errors in classifying a person's levels of intelligence can occur even if the short form IQ estimate is only one point away from the actual IQ (e.g. 69 IQ vs. 70 IQ)." IQ scores on the WISC-III and the K-BIT are "obtained" scores and are best represented by reporting them in confidence bands. In order to investigate the meaning of these "misclassifications" two approaches were taken.

First, given the high reliabilities of the two measures (WISC-III Full Scale = .96, K-BIT = .94) and the correlation between the measures found in this study ($r=.83$), it was possible to compute the magnitude of the difference required for significance. When this was done it was found that a difference of greater than 9 points might be considered a 'significant' difference. Using this information, the WISC-III Full Scale and the K-BIT Composite scores were compared. Of the 613 students tested, 353 (58%) were found to have obtained scores on the tests within 10 points of each other. However, 260 (42%) had scores that were 'significantly' different, with the K-BIT overestimating 209 (34%) students while underestimating 51 (8%).

Second, a 'practical" approach to the misclassification was investigated. Every one of the 350 children with a differing classification was identified. The two classification labels (one from the K-BIT Composite and one from the WISC-III FSIQ) were compared to see if they were at least in adjacent categories (for example: a WISC-III FSIQ classification of "AVERAGE" was compared to the K-BIT classification to see

if the K-BIT was in either the "LOW AVERAGE" or "ABOVE AVERAGE" range.) This allowed us to determine if the misclassification extended beyond a single label. When this was done, 97 children (16%) has scores from the tests that placed them at least two classification labels apart.

Results of this investigation suggest that although the K-BIT is an adequate screening instrument for use in a pre-referral evaluation process, caution must be taken to ensure that the scores obtained from the K-BIT are not used in determining eligibility for special education services. As noted in the K-BIT manual, although it has the same mean and standard deviation as the Wechsler and Kaufman scales, "...it does not imply that the K-BIT may substitute for a comprehensive measure of a child's or adult's intelligence."


Flynn, J. R., (1987) Massive gains in 14 nations: What IQ tests really measure. Psychological Bulletin, 95, 29-51.
Kaufman, A. S. (1990) Assessing Adolescent and Adult Intelligence. Allyn and Bacon.
Weiss, L. G. (1995). WISC-III IQs: New norms raise queries. in Assessment Focus. The Psychological Corporation



C. L. Boyd NCSP is Director of the Psychological Assessment Center for the School Board of Polk County, Florida



**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792387 HITOMETER



# Continuous Performance Tests

**reviewed by Ron Dumont, Anna Tamborra, and Brian Stone**

**A similar review by these authors appeared in the NASP Communiqué, 24, 3, 22-24**

**Link to comparison table**

Three computerized continuous performance tests were reviewed by these authors. The goal of these reviews was to compare the ease of use, computer requirements, normative data, test result, and interpretability. No attempt was made to distinguish which program might "better identify" a sample of ADHD children from a control group. Materials reviewed generally were those that a practitioner would receive when purchasing the software package. Although extensive research may already have been published on CPTs, it was our goal to review only the materials that a school psychologist would receive when purchasing the tests themselves. Admittedly, no attempt has been made to extensively research the background of the tests and their historical use. These reviews should be considered general overviews of the tests and are not meant to be comprehensive in their nature.

## Test of Variables of Attention (T.O.V.A.) (Link to users comments)

The Test of Variables of Attention (T.O.V.A.), is a computerized, 23-minute (11 minutes for 4-5 year olds), non-language based, fixed interval, visual performance test for use in the screening; diagnosis; and monitoring of treatment of children and adults with attention deficits. It was created by Dr. Lawrence Greenberg and is distributed by Universal Attention Disorders, Inc. as well as American Guidance Service (AGS). Cost for this test is $495. This price includes the T.O.V.A. disk, micro switch (button), the T.O.V.A. box (for keeping track of additional tests), two T.O.V.A. videos, an interpretation manual, and an installation manual. The initial cost also allows 5 interpretations. Each additional interpretation costs between $5 and $6 depending on the number you purchase. (Included in packets sent to us a year before doing this review were a number of interesting materials that probably reflect the difference in the professions of those who might use the T.O.V.A.. On one promotional page, under the heading "Benefits of T.O.V.A.", the following were listed: Enhance revenues; Retain patient within doctor's practice; Builds practice/builds referral base; and Reimbursable through major medical/psychological benefits. Also included were two sample letters to insurance companies demonstrating how to bill for using the T.O.V.A. for either a C.N.S. diagnosis or an Organic Brain Syndrome Diagnosis.) Could school districts or school psychologists that use the T.O.V.A. request third party reimbursement?

The manual states three clinical uses of the T.O.V.A.: 1. as a screen for students suspected of having ADHD or learning problems; 2. as a diagnostic tool as part of a multi-disciplinary assessment of children and adults who may have attention deficit; and 3. as an aid in helping to determine the dosage level and to monitor the use of medication over time.

The test itself consists of repeated exposures on the computer screen of two different squares. The squares differ in that one has a 'hole' near

the top (target figure) while the second has a 'hole' near the bottom. The subject is to press the button every time the square with the hole near the top is flashed on the screen. The T.O.V.A. variables include: Errors of omission (inattention) and commission (impulsivity); response time; standard deviations, anticipatory responses, post-commission responses, and multiple responses.

The Test of Variables of Attention (T.O.V.A.) computer program (version 1.3.1) was reviewed using a PowerMac 7100/66, with 16MB of memory. The manual that accompanied the software was for version 1.2. Installation of the T.O.V.A. software itself was flawless. Simply dragging the T.O.V.A. icon to the hard drive installs the program. The problem came when trying to connect the T.O.V.A. button to the modem port. It didn't fit. The configuration of the Mac's serial port had evidently changed from earlier models to the Power Macs and the 8 pin plug provided for the T.O.V.A. would not fit into the serial port. Luckily a toll free number is provided for technical support. After speaking with Andrew Greenberg, we chose to attempted the "low tech" solution of using an exacto knife to do away with the plastic surrounding the pins. When this didn't work, Andrew gladly sent, and we received in 1 day, a micro processing switch that solve the problem. Technical support for computer problems and the availability of people knowledgeable of the T.O.V.A. when questions arose was excellent throughout the reviewing process.

One strange alert box appeared before the correct micro-switch arrived. The alert suggested that we might choose to use the computer mouse button instead of the micro-switch and directs the user to push the "Use mouse" button. There was however no button to click on! This is probably a good thing, since the test measures response time in microseconds and any inaccuracies would greatly affect the interpretation of the T.O.V.A..

What appears to be an error in the computer program was discovered when we entered the age of a child as 7 years old and the computer generated the incorrect form (#6 - Age 4-5 (IF)). Once the form had been set by the computer at #6, it could not be brought back to the correct age form (#1) without creating a new test subject set-up. Examiners not aware of this form change requirement could in fact administer the wrong test to the subject. (Andrew Greenberg reported that this error would be fixed immediately.)

Another caution must be noted. It is possible for the results to differ from the child's actual performance. We found during one administration that when the results were sent to the printer, certain scores (omission errors) were reported when they had not occurred during the test taking. This was possibly caused by a 'powering-down' energy saving system in the printer hardware. To avoid this problem, examiners are cautioned to be sure they are using a printer that is fully on line from start to finish and that examiners remain with, and closely observe the actual performance of each person tested.

The manual provides normative data on 1590 subjects, at 15 different ages separated by sex. Male and female norms are reported separately because, on the average, males have faster reaction times but make significantly more errors of commission (impulsive guessing). The norms clearly show that sustained attention increases with age, levels out at adulthood, and then deteriorates slightly in older adults. The norms are not stratified and little, if any, information is provided about the makeup of these children and adults. No breakdowns for socioeconomic levels, geographic regions, education levels, or race information is provided. There is no evidence in the manual that the normative sample includes (or for that matter, excludes) special education students or children on stimulant medication. Above age 20, there are very few males in the norming tables. For ages above 19 the numbers in the norming sample age groups drops considerably from an average of 168 subjects per group (age 4 to 19) to 36 subjects per group (age 20-80+). At some ages male subjects in the norm sample made no errors, hence there was no variability. Thus, actual standard scores at these points are quite artificial. (In separately published and unpublished information not sent with the test materials, the T.O.V.A. normative group appears to be created from "rolling norms", the continual addition of people to the

sample at varying stages and then recalibrating the averages. An early sample included 775 children aged 6 to 16. These children came from grades 1, 3, 5, 7 and 9 in three Minneapolis, Minnesota, suburban public schools. The children were "mainly middle to upper-middle social class and was predominantly Caucasian (99%). A second sample of 821 children and adults was later added to the original total. These new subjects came from an early education screening project; randomly selected classes in one grade- and one high-school in a rural Minnesota community; volunteer undergraduates in three Minnesota liberal arts colleges; and adults living in six adult community settings. Children in special education classes were excluded from each sampling. The total number of subjects in the norming sample is somewhat confusing. The print-outs from the T.O.V.A. states that the norming base is "of 2000 children and adults." The manual presents data dated 7/94 that includes 1590 children and adults. These are entitled "revised norms" yet are the same as those published in a paper dated 9/92. There is no mention of the remaining 400+ subjects.) Still, the norm sample is impressive for a test not published by a major company.

The primary author of the T.O.V.A., Dr. Lawrence Greenberg, is a psychiatrist, and the concepts of reliability and validity appear to be addressed in a somewhat different fashion than is typical in our field. To it's credit, many differential diagnoses studies are cited where the T.O.V.A. is used (alone, and in conjunction with the Connors Parent Teacher Questionnaire (CPTQ)) to discriminate between children with attention deficit disorder and normals (also children with behavior disorders/and other diagnoses). The T.O.V.A. appears to have good sensitivity and specificity in this regard, particularly when used, as the authors recommend, in conjunction with other instruments. The T.O.V.A. was best at differentiating between attention deficit and normals. Still some normals overlap with some attention deficit disorder children. (There are no studies to show the T.O.V.A. is able to differentiate an attentional disorder from a specific learning disability. One statement made in the promotional materials and restated on the video is that because the T.O.V.A. uses a task that is "non-language based" it can differentiate ADD from learning disorders. We are not sure that that statement is sufficient to prove the point. If this was in fact true, why were the special education students excluded from at least the original normative sampling? It might have been helpful to have tested children identified as having a specific learning disability and then compare those results to the normative group.)

One concurrent validity study with very few subjects looked at the overlap between the T.O.V.A. and the CPTQ. Unfortunately, the authors employed a canonical correlation with 23 subjects and approximately 10 variables (it was unclear exactly how many variables were utilized). This is far too few subjects for such a study, and is therefore uninterpretable.

The authors looked at test-retest data for 97 subjects across ages and found no significant differences between testings "except for commission errors which...improved during the first half of the test from first to second test but not for two subsequent tests." (Manual, p. 2). Interestingly, the authors note that practice effects tend to be reverse of other tests, in that subjects tend to do worse on it, as the novelty of the stimulus wears off. Overall, the authors concept of reliability in the manual refers to what are basically "lie scales", however, these scales appear very useful in telling if the subject is merely responding at random. Psychometric reliability data would be welcome.

More validity studies would be useful, particularly in a divergent/convergent framework (e.g., does reaction time (or any of the measures) as used in the T.O.V.A. correlate with cognitive ability; do they correlate with other observable behaviors, etc.). Correlation between the different T.O.V.A. measures would also be useful. The authors state that they assume that a child 2 standard deviations below the mean on IQ would also be 2 standard deviations below the mean on the T.O.V.A.. Actually, the lower the g-loading of a given T.O.V.A. measure, the more it would tend to regress (be closer to) the mean.

The authors do an excellent job at showing how stimulant therapy affects T.O.V.A. responses. The T.O.V.A. appears particularly useful in being used to establish a baseline, prior to stimulant medication, then used to monitor stimulant medication afterwards. The T.O.V.A.

measures appear very sensitive to stimulant therapy. This finding is quite impressive and certainly bolsters the validity of the T.O.V.A..

The authors take great care to point out the T.O.V.A. is not meant to diagnose attention deficit disorders, but is a good screener, and is useful as a part of a larger battery. They advocate behavioral interventions, possibly in conjunction with stimulant medication.

The T.O.V.A. was easy to load and run. The program worked effortlessly with the minor exceptions noted above. It is purposefully boring, and probably more so for the examiner who must sit patiently through the 23 minute test. Examiners may find themselves leaving the client alone while the test continues, but this seems like a bad idea since the clients behavior during the testing may be important in the interpretation of the results.

The T.O.V.A. looks promising and would make a good tool for further research. The manuals are replete with typos, but perhaps that was a test of our vigilance. The manuals could have benefited from a historical and theoretical perspective, as well. Overall, the test would certainly benefit from more of the typical reliability and validity data, but was impressive in many areas, including differential diagnosis and sensitivity to stimulant medication. It should serve researcher's well, and would be fun to use for those considering masters' thesis and doctoral dissertation work in the area.

**Conners' Continuous Performance Test (CPT)**

The Conners' Continual Performance Test (CPT) is a computerized, 14-minute, visual performance task in which the subject must respond repeatedly to non target figures and then inhibit responding whenever the infrequently presented target figure appears. The test is a "useful attention and learning disorder measure for children, and is sensitive to drug treatment in hyperactive children." The manual states that the program is most useful for children between the ages of 6 and 17. Among the many variables are: Number of Hits, Omission, Commission; Response Time. It was created by Dr. Keith Conners and is distributed by Multi-Health Systems, Inc. as well as The Psychological Corporation. Cost for this test is $495. This price includes the CPT disk, and an interpretation and installation manual. The program offers unlimited administration, scoring and interpretations of the complete "Standard" paradigm. For research purposes, the computer program offers the ability to create customized paradigms with varying letters, presentation time, trials per block, etc.. It must be noted that normative data is only available for the standard paradigm. Anyone using the customized paradigm must do so with the understanding that no normative data is available for any such changes.

The "Standard" test itself requires the subject is to press the appropriate mouse button or the keyboard's spacebar for any letter except the letter X. There are 6 blocks, with 3 sub-blocks each of 20 trials (letters presented whether target or not). For each block, the sub-blocks have different stimulus intervals. These intervals vary between blocks.

The Conners' CPT computer program was reviewed using an IBM computer as well as a Power Mac 7100/66, running Soft Windows with 16MB of memory. Although the program was easily loaded onto the Power Mac, it could not be run under the simulated DOS. A toll free technical support number is available for anyone having difficulty with the program. The first time we called we were put on hold for 30 minutes before the technical support person came on. The next two times we were connected to technical support within a minute. All questions were answered quickly and courteously. Once properly installed on the IBM, the program ran flawlessly.

The manual provides normative data on 1190 subjects, at 8 different age groupings. This sampling is further broken down into two groups:

General population (n=520) and Clinical sample (n=670). Careful reading of the manual indicates that this clinical sample was further broken down to 484 people after 130 subjects were removed for a cross validation study, 46 removed for being "outliers", and 10 more removed because of being on medication. The 484 was finally reduced to 238 subjects comprised of ADD/ADHD and comorbid cases (including ADD/ADHD as one of the diagnoses). Male and female norms are used by the computer program but are not reported separately in the manual. The "general population" and "clinical population" consisted of 51.2% and 75.4% males respectively. No breakdown by age category is offered. (In fact, in the manual, no normative score data is given with the exception of that stated above). The norms are not stratified and little, if any, information is provided about the makeup of these children and adults. Very little information regarding socioeconomic levels, geographic regions, education levels, or race information is provided. It is noted that data for the general population came from 5 states and "Southern Ontario."

The program provides data as both raw scores, T scores, percentiles, and descriptive classifications (e.g.., Within Average range, Mildly atypical, etc.). Reports are available on screen, as a print out, and as an ASCII file saved to disk.

The concepts of reliability and validity were not addressed thoroughly in the manual. It appears from reading the extensive annotated bibliography that some studies may have been carried out by independent researchers. However, with only the manual to rely on, we were left with many questions regarding these issues.

The major validity issue addressed in the manual looked at the ability of the CPT to discriminate between children with attention deficit disorder, "normals" (includes children with behavior disorders/and other diagnoses), and a comorbid group (children with dual diagnoses of ADHD and other disorders).

The CPT appeared to discriminate well, typically having the poorest mean score in the pure ADHD group, a somewhat better mean score in the comorbid group, and the "best" mean score in the "other" group, for the majority of variables. Unfortunately, the standard deviations were not listed, so the degree of overlap between groups on these variables is unknown. Another statistical technique, such as discriminant analysis would have been nice. Also problematic would be the existence of subtypes of ADHD within the ADHD sample. Perhaps the greatest display of validity is the letter of support issued in Russell Barkley's newsletter that states the CPT is very much in line with current theory compared to many other instruments on the market (1993, June).

The manual seemed more concerned with history and theory than reliability and validity issues. The admissions in the manual were well appreciated, including the variability in sustained attention across times with the same subject, and the idea that, like IQ, there are many reasons for poor scores.

More research is needed on the stability of the many variables this test offers. Also needed is information regarding the independence of these variables (are they highly correlated with each other? What other measures do they correlate with?). Some of the independent research listed addressed these questions, but often the short abstracts of the studies listed were far too scanty to cull such information from.

However, kudos to the publisher for compiling the reference bibliography with abstracts (the little information contained was tantalizing and should send many buyers to their respective research libraries.)

To the author's credit, an excellent job is done at showing how stimulant therapy affects CPT responses. The author also takes great care to

point out throughout the manual that the CPT is not meant to diagnose attention deficit disorders by itself, and is useful as a part of a larger battery.

**Intermediate Visual and Auditory Continuous Performance Test (IVA)**

The Intermediate Visual and Auditory Continual Performance Test (CPT) is a computerized, 13-minute, visual and auditory performance task in which the subject must click the mouse only when he or she sees or hears the number 1 and not click when he or she hears or sees the number 2. The test is designed to assess two major factors: response control and Attention. In addition, the IVA provides "an objective measure of fine motor hyperactivity." The manual states that the program is useful for persons between the ages of 5 and 90+. Among the many variables are six core quotients and 22 subscales. It was created by Drs. Joseph Sanford and Ann Turner and is distributed by BrainTrain. Cost for this test varies. A limited use kit (25 administrations) costs $598. This price includes the IVA disk (IBM 3.5 or IBM 5.25), and an interpretation and installation manual. Disks with an additional 25 tests may be purchased at a cost of $75. Users also have the option of purchasing an "Unlimited Use Version" for $1495.

The IVA CPT computer program requires an IBM computer with DOS 5.0 or later; 1 MB RAM/2 MB harddrive; a graphic monitor; serial mouse (Microsoft recommended); Creative labs Soundblaster card; Headphones or external speakers. (A toll free technical support number is available for anyone having questions with the program or the interpretation.) These requirements caused the most difficulty for these reviewers. In order to properly run the program, we had to find a computer that met each of the requirements, the most important being the Soundblaster card, the Microsoft mouse, and the headphones. Those well versed in IBM computers may feel right at home with this product, but these reviewers struggled for over an hour trying to get the mouse driver configured and the sound card and driver running. The installation of the program itself was not difficult. Step by step directions are provided in the well written manual. One hopes that the program can be re-written for a Macintosh since those computers come with voice capability and speakers built in.

The program uses normative data from 781 subjects (423 female, 358 male), at 10 different age groupings. No breakdown by age category is offered in the manual. There is no evidence that the norms are stratified and little, if any, demographic information is provided about these subjects. No information regarding socioeconomic levels, geographic regions, education levels, or race is provided. It is noted that the groups were comprised only of persons "who do not report any attention, learning, neurological or psychological problems." The normative data file, contained on the program disk, was easily read by us using a Macintosh computer. The 10 age groups averaged approximately 42 female (range 15 at age 55+ to 75 at age 7-8) and 36 males (range 17 at age 45 - 54 to 68 at age 7-8). Age groupings are: 2 years (5 - 10), 3 years (11 - 13), 4 years (14 - 17), 7 years (18 - 24) and 10 years (25 - 55+).

The program provides data as both Quotient scores (mean 100, SD 15), percentiles. Graphs also are used to represent the results. The interpretation section of the manual is easy to read yet quite complex. The 6 quotient scores plus 22 subscales offer a large number of decisions and comparisons. The manual presents 17 pages of description and definition for each scale and 34 pages of interpretive suggestions. Included is a 21 step "procedural guidelines" for interpreting the IVA. The program offers 3 "Validity" scales used to confirm or refute the IVA results. Reports are available both on screen and as printouts. Data is stored and available on disk for retesting and comparisons.

The packet we reviewed contained 5 unpublished studies (presented at the 1995 APA conference). These studies address: normative, reliability, and validity data, differences in auditory and visual processing, and finally developmental age and sex differences on the IVA.

The extremely well-written manual was readable and informative. It addressed both reliability (stability?) and validity issues. It also reported (admitted) the less than stellar test/retest correlations across some variables (ranged from .37 to .75 for the composite variables). Even though the studies were only APA conference presentations, the authors have attempted to look at the important issues. An important question as yet unanswered by the materials included was to what extent did the auditory and visual variables correlate - how separate are they? Are they as highly correlated as the Wechsler verbal and performance scales? If so, are they subject to the same argument proposed by MacMann and Barnett who suggest that the Wechsler scales are so highly correlated as to render them similar measures of the same construct (for a review, see Kaufman, 1994).

Particularly impressive were the three validity scales, which ensure scores in ADHD ranges come from ADHD behaviors and not motor problems, fatigue, or random answering. The manual also acknowledged the relationship between IQ and sustained attention, and suggested IQ scores of 120 and above would be well-served by comparison to the next age norm table up. The authors also acknowledged the issue of subtypes of ADHD (inattentive, impulsive, "mixed," and other). The authors addressed the issues head on, and asked the best questions. While all three tests addressed reliability and validity to some degree, the IVAs authors did the best job at asking the right questions. While all three tests are really in the beginning stages of compiling research data on reliability and validity, the IVAs authors are headed in the most compelling direction.

**Test takers point of view**

Each of the CPTs was administered to one or more of these reviewers to assess ease of use from the test takers point of view. Since the tests varied in length from 13 to 23 minutes, plus some additional time for practice testing, we found that our attention varied as the tests extended in time. Early in the testing sessions, we found ourselves being very cautious and completely focused on the screen, but as the tests continued on in time, it seemed more and more difficult to maintain our focus on the stimuli from the computer. The speed of some of the tests' stimuli presentations was so rapid that we found ourselves almost afraid to blink. This led to our eyes becoming tired, and to a heightened sense of anxiety. (Because the tests are standardized, we assume that those included in the norming sample probably felt some similar feelings, and the normative scores adjust for such feelings.) Testing was done in a fairly sterile room, but we did not attempt to 'sanitize' the room completely. There were some materials on the walls and in shelves around the computer. We found that even these few things became very tempting distractions during the testing. Not only did we find ourselves easily distracted by these visual materials, we found ourselves being drawn to and distracted by sounds outside the testing room. Each manual provided instructions to the testers about how to create a positive testing environment, and we strongly reinforce the need to follow these instructions. Any extraneous material may have the potential to interfere with performance. Because the examiner needs to be present during each of these CPTs, the examiners must assure that they do not become a distraction themselves by unnecessarily moving about or making any noise. (This may become somewhat difficult, especially after sitting through a number of these admittedly 'boring' tests.) One final caution we learned by taking the tests was the need to give the directions exactly as they are presented in the manual. For example, one of us took a test without having the instructions read verbatim from the manual, and without any emphasis placed on the direction to do the task "As fast as you can." Interestingly, the resulting printout recommended further assessment because of the suspicion of an attention disorder.

**CPT Bottom Line**

Choosing between the different tests will depend on many individual factors. The three CPTs reviewed in this issue of the Communiqué each offers something unique to the examiner and examinee. The T.O.V.A. uses a design (square), Conners' letters, and the IVA numbers (1 and

2). The IVA is the only CPT to offer both auditory and visual procedures. If cost is a factor, the Conners' is the least expensive while the IVA the most expensive. If the computer system is an issue, the T.O.V.A. is the only CPT that runs on a Macintosh, while all three offer versions for IBM based machines. The T.O.V.A. and the Conners' require the least amount of "extra" hardware. Normative data for the CPTs was largest for the T.O.V.A. although none of the test provided enough demographic information about the subjects to make informed judgments about the suitability of the data. If time is a factor, the Conners' and the IVA were the shortest tests. Ease of use was comparable for each of the tests. Support for the programs by way of toll free telephone numbers is provided by each system.

Our experience using these three programs was generally very positive. We stress however that the programs are simply one tool to be used in a multi-dimensional assessment. Each test product included clear warnings about not basing diagnosis on the single instrument or result. We whole-heartedly endorse this caution, especially given the vast differences between computers, computer systems, and the few 'kinks' we discovered during our limited reviews.

The editors of the Communiqué would like to thank each of the three companies for providing the programs for review.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792388 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Continuous Performance Tasks Compared

| Test | Test of Variables of Attention (T.O.V.A.) | Conners' Continuous Performance Test (CPT) | Intermediate Visual and Auditory CPT (IVA) |
|---|---|---|---|
| **Publisher** | Universal Attention Disorders, Inc. 800-729-2886 | Multi-Health Systems Inc. 800-456-3003 | BrainTrain 800-822-0538 |
| **Author** | Dr. Lawrence Greenberg | Dr. C. Keith Conners | Drs. Joseph Sandford & Ann Turner |
| **Materials** | T.O.V.A. program; T.O.V.A. button; EVE3 hardware; manual; video | CPT program, manual | IVA program, manuals |
| **Cost** | $495*<br><br>*additional charge per interpretation after the first five | $495*<br><br>*unlimited use | $598*<br><br>*additional charge per interpretation after the first twenty-five. Unlimited use software available at $1495 |
| **Computer** | IBM: DOS 3.0 or later; 380K RAM; Hard drive; 720 disk capacity; VGA, CGA, or EGA graphic capability; Parallel port<br><br>Macintosh: Any Mac later than the Macintosh SE; any Mac-compatible printer; System 6.0.7 or greater (System 7 requires at least 4MB of RAM) | IBM: DOS 3.3 or later; 512K RAM; Hard drive; Monochrome or Color monitor; mouse (optional); printer (optional) | IBM: DOS 5.0 or later; 1 MB RAM/2 MB hard drive; VGA, CGA, or EGA graphic monitor; Serial mouse (Microsoft recommended); Creative labs Soundblaster card; Headphones or external speakers; Printer (optional) |
| **Norms** | 4 to 80+<br><br>1590 subjects at 15 ages separated by sex. Norms are presented in 2-year intervals from 4 to 19, and then in 10-year intervals from 20 to 80+. | 4 to 18+<br><br>520 subjects at 8 ages. Norms are presented in 2-year intervals from 4 to 17, and then as one group 18+. | 5 to 90+<br><br>781 subjects at 10 ages separated by sex. Norms are presented in differing year intervals ranging from 2 to 7 years for ages 5 to 24, to 10-year intervals from ages 25 to 55+. |

| | | | |
|---|---|---|---|
| **Description** | A 23-minute, non-language based, fixed interval, visual performance test for use in the diagnosis and monitoring of treatment of children and adults with attention deficits. | A 14-minute visual performance task in which the subject must respond repeatedly to nontarget figures and then inhibit responding whenever the infrequently presented target figure appears. | A 13 minute, auditory and visual performance task in which the subject is required to click the mouse button only when he or she sees or hears a 1 and not click when he or she sees or hears a 2. |
| **Major Variables** | Errors of Omission (Inattention) and Commission (Impulsivity); Response Time; and Variability | Errors of Omission and Commission; Reaction Time | Six composite scores and 22 raw scales |
| **Report** | On screen and printout | On screen and printout | On screen and printout |

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792389 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# SENSITIVITY, SPECIFICITY, AND POSITIVE AND NEGATIVE PREDICTIVE POWER FOR CONTINUOUS PERFORMANCE TESTS

**Ron P. Dumont, Ed.D., John Willis, Ed.D., Casey Stevens, M.A., C.A.S.**

The Gordon Diagnostic System is another in the array of computerized CPTs (Continuous Performance Tests) that have been used to help diagnose attentional difficulties. It, like many of the CPTs, lacks strong validity. I do not deny that CPTs can be useful in a comprehensive evaluation of children suspected of attentional problems, but they are not 'proof' in and of themselves. There are a number of important confounds that could impact upon the score obtained on this and other computerized measures.

Reviews of research on computerized continuous performance tasks have generally been favorable, and they are seen as playing a role, albeit limited, in the evaluation of attention disorders. Barkley and Grodzinksi (1994), for instance, evaluated the utility of neuropsychological measures, including continuous performance tests (CPTs) for distinguishing children with ADHD from normal controls and children with learning disabilities. They found CPT measures among the most useful of the assessment procedures investigated. Nonetheless, they noted that positive but not negative CPT findings can have diagnostic utility. Thus, while poor performance on a CPT measure was indicative of an attention disorder, good performance did not necessarily rule out attention disorders. These results have also been replicated by Matier-Sharma and colleagues (Matier-Sharma et. al., 1995).

One recent study, Wherry et al., 1993, Psychology in the Schools, investigated the validity of the GDS and the results were fairly poor. These authors stated that "The results failed to demonstrate the discriminant validity of any GDS score regardless of the behavior rating used." As Barkley and others (1994) have noted, in order for a test to be diagnostically useful, it must be able to not only identify the children with ADHD, but it must also accurately identify children without ADHD. One very important issue regarding the typical validity studies is their use of already identified clients. This does provide some aspect of validity but it is also necessary to investigate the sensitivity and specificity of the measures (something typically lacking).

Ellwood (1993) discusses parameters that can be used to examine a test's diagnostic usefulness. Test specific parameters include sensitivity, or the proportion of individuals with a disorder that exhibit the sign (i.e., the proportion of children with ADHD who receive scores within the abnormal range) and specificity, or the proportion of individuals without a disorder that do not exhibit the sign (i.e., the proportion of controls who receive scores within the normal range). These two parameters are calculated in the research setting by first knowing the diagnosis of the children (through test-independent criteria) and noting how they perform on the test of interest. However, as Ellwood (1993) points out, this is the opposite of the way an evaluator uses a test. The evaluator starts with the test score and attempts to determine the child's diagnosis. In order to judge the usefulness of a test for this purpose, the evaluator will need to look at a test's sensitivity and specificity in light of the disorder's base rate in their referral population.

For example, if a test was used as a screening measure on a population of 1000 children in which 4% (40) of the children have ADHD, and

that test gives an abnormal score for 90% of the children with ADHD (i.e., sensitivity) and gives a normal score for 90% of the children without ADHD (specificity), the following diagnostic properties result.

**Calculation of Sensitivity, Specificity, PPP, and NPP**

| | ADHD | Control | |
|---|---|---|---|
| | **a** | **b** | |
| Abnormal Score | 36 | 96 | 132 |
| | **c** | **d** | |
| Normal Score | 4 | 864 | 868 |
| | 40 | 960 | 1000 |

Sensitivity = a/a+c = .90

Specificity = d/b+d = .90

PPP = a/a+b = .27

NPP = d/d+c = .99

Using this table, one can calculate Positive Predictive Power (PPP), or the chances that a child who receives an abnormal test score actually has ADHD. PPP = a/a+b = 36/132 = 0.27. A test with 90% sensitivity and specificity has restricted usefulness as a diagnostic tool if it is used on a population with a 4% base rate of the disorder because if the child receives an abnormal score, (s)he is still much more likely to be a control than a child with ADHD.

The issue of PPP and NPP was from an article we wrote on the use of the Mesulam CPT (a paper and pencil CPT) that takes about 3 minutes and can be used in an entire class. We found the PPP and NPP to be similar or better than the Computerized tests.

Barkley, R. A. and Grodzinksi, G. M. (1994). Are tests of frontal lobe functions useful in the diagnosis of Attention Deficit Disorders? *The Clinical Neurologist*, *8*, 121-139.

Ellwood, R.W. (1993). Clinical discriminations and neuropsychological tests: An appeal to Bayes' theorem. *The Clinical Neuropsychologist, 7*, 224-233.

Matier-Sharma, K., Perachio, N., Newcorn, J.H., Sharma, V., & Halperin, J. M. (1995). Differential diagnosis of ADHD: Are objective measures of attention, impulsivity, and activity level helpful? *Child Neuropsychology, 1*, 118-127.

Wherry, J. N., Paal, N., Jolly, J. B., Balkozar, A., Holloway, C., Everett, B., & Vaught, L. (1993). Concurrent and discriminant validity of the Gordon Diagnostic System: A preliminary study. *Psychology in the Schools, 1*, 29-36.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792390 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Software Review: A Comparison of Five Interpreter/Report Writers

Ron Dumont - Casey Stevens - Jon Short

---



---

## Introduction

These authors reviewed 6 commercial software packages currently available for use in creating psycho educational assessment reports. All programs were provided by the individual software companies with no stipulations placed on the authors' comments. (One company requested a signed contract that allowed for prepublication editorial input and that software package was not reviewed). IBM compatible programs (both DOS and Windows based) were reviewed using a Digital Venturis 466, while Macintosh programs were reviewed on a Power Macintosh 7100 and a Powerbook 5300cs. The goals of these reviews were to compare and contrast the programs' ease of use, user friendliness, information provided, and accuracy of results.

For each of the programs, installation on the appropriate machines was simple and straight forward. Those familiar with the particular computer should find no difficulty loading and running either version. Computer neophytes will find the installation instructions in the manuals helpful. Once loaded and running, all programs offer both mouse and/or keystroke navigation.

Each program reviewed acted, to greatly varying degrees, as a report generator by placing inputted scores (raw or scaled scores depending on the program) and user chosen descriptors into its own report format. The resulting narrative is editable either within the program itself or after the program transfers the results to your word processor. Most reports included areas for background information, test results, differing levels and ways of interpretation, a section for recommendations (program or user supplied) and some form of tables that include all relevant scores. Most programs provided a way to input scores from achievement tests and to then determine if there was an ability/achievement discrepancy. Differing options for making this determination were available. Most programs print a 5 to 10 page report. No program contained a built-in spell checker. Each program offered varying degrees of error checking of inputted scores. Some programs were found to have serious errors that can effect the interpretations generated. Anyone using a computer program must be mindful of the legal and ethical obligations that accompany their use.

WISC-III Writer: The Interpretive Software System. The Psychological Corporation, $295. Available for Macintosh and IBM (DOS and Windows based).

This program allows for the scoring and interpreting of the WISC-III and the WIAT. Users are able to input either raw scores or scaled scores from the tests and the program converts these scores into appropriate standard scores and IQs. Because this data is copyrighted by the Psychological Corporation, this is the only commercial program available that allows such a conversion of scores. Extensive error checking is available. Scores entered by accident that do not match the range expected are automatically flagged with a warning dialog. Once the data are entered, they may be saved to a database and sorted by a number of options. If the user has graphic capability, this program also allows for the inclusion of a child's picture. Once scored, the computer will generate varying reports: a parent report; an interpretive report; or a set of Tables and Graphs. Besides the scoring and interpreting of the Wechsler material, the extensive report writing nature of the program comes from 20 selectable screens that provide the user with check boxes and/or radio button choices. Areas covered include: Test session behaviors, Referral questions, Previous evaluations, Medical results, etc.. The program will tailor the printout based on the information checked in these windows. The choices made by the user are checked for contradictions and the program guards against them (in the development section, if a user checks "was born with no apparent complications" and then checks "was born prematurely", the program automatically un-selects the first choice. The report generated is fully editable on screen before printing. Also included in the program is a fairly detailed set of "ReportClips". These are prewritten statements that can be copied and pasted directly into the current report. These ReportClips cover 11 broad areas including: Academic, Intellectual, Recommendations, and Speech and Communications. The program also allows the user to create and save new ReportClips for later use. This may be especially helpful for users who wish to save pre-written statements that are used continually in different reports. Users have the ability to save the completed reports as report files (separate from the database), as well as the ability to copy the entire report into a word processing program for additional editing and formatting. This was useful since the program does not contain a spell checker. It should be noted that when the reports were spell checked, no errors were found in any of the statements generated by the program.

The program offers the user many options for configuring the interpretive report and analysis. All options are easily changed with a click of the mouse. A full set of options for Table printouts; Ability estimates (IQ vs. Indexes); Ability achievement discrepancy determination (simple difference vs. Predicted difference); Significance levels for confidence bands and VIQ-PIQ differences; Headers and Footers are included. Users are also given the option to change the classification labels attached to certain IQ ranges (i.e., 80-89 = Low average could be changed to 80-89 = Below Average).

The manual was useful for understanding how to use the program but especially so for understanding decisions made by the program in the development of the interpretive report. Chapter 2, Interpretive Rationale, was an easy to understand description of the decision making algorithms used by the program. The rationale is similar to, yet different from, that proposed by Alan Kaufman in Intelligent Testing With the WISC-III. It should be noted that the Macintosh and DOS versions of this program do not allow the option of substituting Symbol Search for Coding (a la Kaufman) but the Windows version does.

Inputting of scores is done on separate WISC-III and WIAT pages. If the dates between testing exceed a certain time, the program warns the user in the printout about the validity of interpreting such results. The computer converts the inputted scores (raw or scaled) into scaled scores and percentile ranking. The computer generates a prorated Verbal and/or Performance IQ after 4 subtests from the appropriate scale are entered. (Will this result in people creating "short form" IQs without understanding the technical properties of them?) In order to see interpretive information concerning indices and subtest score strengths and/or weaknesses, users must create a report. The input screens are

not interpretive. Within the report, subtest strengths and weaknesses are developed always using the separate Verbal and Performance test means. Critical values used for developing subtest strengths and weaknesses are from the average of all ages in the standardization sample. If the program chooses to interpret the VCI or the POI, it does not redo the statistical analysis to report strengths and weaknesses within those indexes. The printout reports all scores in the body of the narrative as standard scores, percentiles, and confidence ranges. The descriptive categories given are for the obtained score only, (i.e., A child obtaining a FS IQ of 89 would be classified with the words "Low Average" although the range would be Low Average to Average). The option to compute "Shared-Ability Composites" is available. Groupings hypothesized by Bannatyne, Horn, Dean, Kaufman, and Prifitera & Dersh are printed as standard score conversions and identified as being a strength or a weakness. Instructions and cautions about their use are included in the manual.

An extremely useful feature of this program is the option to generate a "Clinical Review" before printing out the final report. These reviews offer the interpreter rationale for the decisions made by the program as well as additional information for the evaluator (i.e., base rate for obtained V-P difference) not included in the final report.

Printing the report was easily accomplished. Options for printing the entire report or individual pages of it are given in the print alert window.

Pros: Only program that allows the inputting of raw scores. Allows for extensive error checking. Provides analysis and interpretation of both the WISC-III and the WIAT. Extensive options for individualizing the report. Contradictory statements are guarded against. It is the only commercial program currently available for a Mac.

Cons: For those wanting to substitute Symbol Search for Coding, only the Windows version has that option. The ability to change Wechsler classification categories and still be able to print out a report with the Psychological Corporation's name on it seems dangerous. Limited to WISC-III and WIAT (although since it is a totally editable text writer, the user can directly add on to the report before printing).

(Top of page)

## Psych Report Writer Version 4.1: Psych Support Systems, $199, Available for IBM.

The Psych Report Writer is a user-friendly program which attempts to interpret results from a wide assortment of commonly used tests, including the Wechsler Scales (WPPSI-R, WISC-III, WISC-R, WAIS-R), the Kaufman tests (K-ABC, K-BIT, KAIT), and the Stanford-Binet Fourth Edition. The program also has the capability of "hooking up" to the WJ-R Cognitive and Achievement tests through the separate Compuscore program. The Psych Report Writer also has the capability of incorporating into the report results from visual motor tests, the Goodenough-Harris Draw-A-Man test, and adaptive and independent behavior scales. This information, along with test session behavioral observations can be recorded on the Psych Report Writer's Computer Data Sheet. These sheets are then used to transfer the obtained scores to the computer at a convenient time. (While it seems a helpful addition to the software package, the data sheet was found to be somewhat incomplete. For example, the WISC-III Processing Speed Index is missing, the WJ-R Cognitive Test and the WIAT are absent from the page, and only 7 out of 21 subtests are listed under the WJ-R Achievement heading. A spelling error was found under the K-ABC Achievement Test heading, with a subtest listed as "Reading/Understading").

The Psych Report Writer provides options for determining ability/achievement discrepancies using either the "Federal LD Discrepancy Formula" (achievement scores greater than 1.5 standard deviations below the IQ scores are identified as significant), or the "Indiana LD

Discrepancy Model" (a regression method which uses the correlations between the IQ and Achievement tests to predict achievement scores). With either formula, the user is able to change the criterion level each time the program is run.

Users will find the program easy to install, and the manual quite helpful. Help options are also available on-screen. The program was easy to navigate with the tab button or the spacebar, however, the mouse cannot be used to simply "click" on the area(s) of interest. The user can review/edit the report before printing, and can set preferences for the header.

The Psych Report Writer does not contain any of the normative data for the tests it interprets and therefore cannot compute the exact IQs (Verbal, Performance, Full Scale, Crystallized, Fluid, Composite, BCA), Index Factor Scores, or Cognitive Ability Clusters. The program provides little or no error-checking for the scores inputted. If an examiner accidentally inputs an incorrect score, the program simply accepts the number and bases interpretation on it. For example, upon entering high scaled scores on the WAIS-R, the Psych Report Writer included in the report indications of Superior to Very Superior abilities on Perceptual Organization and Verbal Comprehension Factors, and yet reported that the subject would "be expected to perform at a level which is significantly lower than same aged peers." It's obvious that these types of results conflict, and the reason for this is that the 'bottom line' recommendation was based directly on the Full Scale IQ, which was entered as zero. Also interesting was that this subject's corresponding classification was listed as "Profoundly Retarded", a term which is not among Wechsler's classifications of intellectual ability.

A number of major errors in interpretation were noted by these reviewers on the Kaufman Adolescent and Adult Intelligence Test (KAIT). The scores inputted are reported at 90% confidence, with the confidence interval for the Composite Score given as a band of ± 8 points, while the KAIT manual provides 90% confidence bands of ± 4 points. More disconcerting was the error related to the development of strengths and weaknesses. The manual for the KAIT makes it clear that the means used for determining significant subtest strengths and weaknesses are to be derived from the sum of 3 Crystallized and 3 Fluid subtests. This is true whether or not the supplemental subtests are given. As stated in the KAIT manual, "The Crystallized and Fluid IQs are based on the three Core subtests whether or not the Expanded Battery is administered" (page 42). During our review, the program consistently gave erroneous means, and thus erroneous strengths and weaknesses, for the cases we examined. Additionally, in one case, the program reported a subtest as a "relative strength" even though it happened to be the lowest score entered. The KAIT interpretation also provides a list of "shared abilities." These did not conform to those in the KAIT manual, occasionally leaving out a subtest that made up the grouping, and consistently creating means that did not match the subtests listed. Another problem concerned the interpretation of significant differences between immediate and delayed memory scaled scores for both Rebus and Auditory Learning. While a discrepancy of four points between the scores was seen as significant at the .01 level, and a discrepancy of three points was seen as significant at the .05 level, significant differences were not reported when the delayed recall subtest score was higher than the immediate recall subtest score. In one case there was a 7 point difference between the relevant scores, yet the program noted that this "was not significant at the .05 confidence level." No rationale is given for this uni-directional mode of significance reporting. One further problem was that users are able to enter scores well over any reasonable limit (300), and even worse, are able to enter such inflated IQs along with below average subtest scaled scores. (Users of this program must check carefully all interpretive findings. These reviewers found serious errors on the KAIT interpretations but did not analyze every possible output and test available from this program).

Again, users are cautioned to examine these reports thoroughly for errors. While no spelling errors were found on the printouts, they were found on-screen ("WAIS-R Age Corected Scaled Scores), and in the manual ("SELECT INTELLEIGENCE TEST", Figure 2.4f). Ultimately the user must be held responsible for the accuracy and spelling on these reports.

Further cautions: In determining discrepancies between IQ and Achievement Tests, the examiner is unable to change the level of statistical significance used (this program is set for 90% confidence). The program comes packaged with 82 "Remediation files", although the extensiveness and current applicability of each is questionable. For instance, remediations are available for students who obtain low scores on any of the WISC-III subtests (except Symbol Search). It is unclear why specialized service is required for children with single, low subtest scores, and even more unclear as to why professionals treating children with low Block Design scores would be recommended to use the following remediation: "more time should be used when introducing new materials to a child with this disability" (bold added). In this same remediation file, practitioners are also cautioned that "such a child tends to learn piecemeal". Furthermore, in a remediation file designed to address poor anger control in students, judgmental and even possibly sexist remarks were found. **"When the inability to control anger is not rooted in organic factors, it stems, of course, from family malpractice. Two aspects of the mother's behavior must be watched:..."** (bold added). No mention of the father's role is made in this remediation report.

(Top of page)

## Kaufman WISC-III Integrated Interpretive System (K-WIIS), version 1. Psychological Assessment Resources, Inc. $425.00

The WISC-III Integrated Interpretive System (K-WIIS) was developed in conjunction with Alan and Nadeen Kaufman and follows closely the interpretive methods outlined in "Intelligent Testing with the WISC-III". The software package comes with computer disks, a manual, and 3 separate checklists which can be used in conjunction with the program. The checklists (Background Information, Physical Observations, and Behaviors Observed During Administration of the WISC-III Subtests) allow the examiner to record basic information during the testing session that will later be transferred into the report. The computer screens mimic the checklists and make entry very simple. Within the program, it is possible to check contradictory observations. (Checking that the mother reported no health problems while pregnant will not prevent one from also checking that she had diabetes, high blood pressure, and toxemia). While running the program, on screen help is accessible at any time and supported by the trouble shooting section in the manual. There is also a 1-800 number listed in the manual for technical support which is available during business hours. The program was user friendly for all levels of computer users. Those unfamiliar with an IBM DOS based program will still have no real problem using the program. Navigation through the program was accomplished easily by either using the mouse or key strokes.

Although the basic function for this program is to interpret the WISC-III results, users are able to enter up to 16 achievement cluster/subtest scores and to compute discrepancies between them and the WISC-III IQs and Indexes. Any achievement score entered must be expressed as scores with a mean of 100 and a standard deviation of ±15. The program will determine both simple and regression based ability/achievement differences. User must supply the reliability coefficient value for the test they wish compared. Information regarding ability/achievement discrepancies is not given in the body of the report, only in the tables appended to the narrative report. Users can set program preferences for confidence intervals, significance levels, and for substituting Digit Span for Arithmetic and/or Symbol Search for Coding. K-WIIS produces a narrative report based on user entered observations, scaled scores, IQ scores, and achievement scores. Recommendations are supplied by both the user and the program. A graph of the individuals reported scores is available, but must be chosen by the user. There is limited error checking (certain ranges for scaled scores and IQs). The program assumes that the user is error free when entering the data. For example, it allowed a Full Scale IQ of 42, a Performance IQ of 90, and a Verbal IQ of 78 to be entered without warning. This could happen by striking the wrong key on the key pad, resulting in an improper report. Text files of commonly used information can be save in a "Library" file and used in subsequent reports by simply copy and pasting the items. Headers and Footers for the report can be altered to suit the users needs. No spell checking abilities are available within the program itself (a few errors were found during this review) but an option does allow the user

to save the report in a specified word processing format and edit the report there.

Pros: Allows for the option of substituting Symbol Search for Coding and Digit Span for Arithmetic. Extensive choices for individualizing the report. Review of the final report can be accomplished on screen before printing and can be saved as a text file to be opened by word processing software. Software package comes with hard copies of the checklists which can filled out prior to running the program. Can be run from DOS or Windows.

Cons: Lack of extensive error checking. Is limited only to the WISC-III. The Mazes subtest is not included at all in the program. No achievement interpretation or narrative in the report.

(Top of page)

QuickWriter Version 1.1d, Ewing Solutions, $249, IBM 386 or newer, Windows 95, Windows 3.1

QuickWriter was easily loaded onto the computer and ran well. Without having read the manual first, we were able to quickly and easily create reports. This report writing program provides a comprehensive psychological report. The reports are generated by choosing options from the "Main Menu" window. This window provides 12 choice buttons, each representing some aspect of the resulting report. After entering a new student's background information or selecting a previously saved case, the user navigates the program by simply clicking the mouse on any of the buttons (keystroke navigation is also available but less desirable). The navigation buttons take the user to screens that offer from one to 6 tabbed sublevels. For example, clicking on the "TEST MENU" button brings the user to a screen with 4 tabbed choices: Test Session Observation (with 6 more submenus), Intellectual (6 submenus), Academic-Adaptive (7 submenus), and Developmental-Visual Motor-Other (6 submenus). Each choice provides numerous check boxes (for multiple statement choices) and/or radio buttons (for choosing mutually exclusive options). Many of the screens also include an active text field in which the user can enter their own statements or comments. This was an extremely attractive option for this program. The program does not allow for on screen editing or viewing of the final report. Reports are either saved to disk, printed, or automatically exported to a word processing program the user specifies.

The QuickWriter provides 3 options for determining ability/achievement discrepancies: a "Standard formula" simple difference method (ability/achievement difference scores greater than a user entered cut-off are identified as significant), a "Federal formula" (a z-score transformation with user defined discrepancy requirements), and the "Arizona formula ( regression method which uses the correlations between the IQ and Achievement tests to predict achievement scores).

No determination of a child's subtest strength or weakness is made by the program. Most interpretive statements simply describe the global scores obtained. More interpretive statements were generated for the WISC-III than any of the other tests available in the report.

We did note one major error and a number of minor problems/errors/discrepancies in the program. The major error involved the interpretation of the WISC-III scores. A statement about the validity of the measured cognitive ability is reported by determining if there is a significant difference between the Verbal Comprehension (VCI) and Freedom from Distractibility (FDI) Indexes. It was found that instead of comparing the score on the FDI to the VCI, the program incorrectly compares the Processing Speed index to the VCI. There are a number of times in the program where, if using the tab key, certain choices are skipped. In the Intervention window, an entire column is missed when using the tab key. A number of the allowable multichoice statements seemed contradictory and might have been better grouped together with radio buttons.

For example, a child's report can conceivably state "Ron is a male with blond, brown, black, red, auburn, fair-haired and dyed hair and blue eyes." A number of spelling errors were noted, most disconcerting was the spelling of Wechsler as Weschler. If one chooses the WAIS-R, there is no instruction about the use of Age Scaled Scores vs. Scaled Scores. The WAIS-R also allows for entering a score on the MAZES subtest! Missing from the Intellectual assessment choices was the Woodcock-Johnson-Cognitive. There was little error checking of the scores inputted. Subtest scores over 19 and IQ scores up to 999 were allowed.

The manual was not particularly helpful for this program. On the one hand, the program is so user friendly there does not appear to be much need for the manual. The manual might better serve if it included more detailed discussion about the decision making processes used in the interpretive narratives.

(Top of page)

## Educational Applications of the WISC-III (EAW3) V2.003, Western Psychological Services, $249, IBM with Windows 3.x

This particular program differs the most from the others reviewed because it does not generate 'typical' narrative reports. The program's focus is to "complement the detailed discussion of WISC-III results presented in "Educational Applications of the WISC-III: A Handbook of Interpretive Strategies and Remedial Recommendations" authored by Charles Nicholson and Charles Alcorn. Installation was simple and flawless. A manual is provided on the disk so that a person wishing may either read it on screen or print it out for reading. The manual explains the basics necessary for installing and using the program but does not provide any explanations for decisions made within the program. Users wishing this information must purchase the "Handbook." Use of the tab key to navigate from entry to entry was easy, and the mouse can be used to point and click into an entry field. Data for the 3 IQs, 4 Indexes, and 13 subtests of the WISC-III are entered as standard or scaled scores. If an achievement test was administered the results of that test can be entered. Although the program has options for choosing 1 of 7 Achievement tests (including the misnamed WIAT-R), the entries for achievement scores do not differ. Users may enter only Reading, Math, and Spelling scores. The program reports all entered IQ and Index scores as single values with no confidence intervals. Comparisons between IQ and academic achievement display standard score differences for VIQ and PIQ only. The program also provides the "Estimated Mental Age and Expected Grade Equivalent (GE) Achievement Level". A theoretical GE Achievement Level at Age 16 is also reported.

The completed "Professional Report", unlike all others reviewed, consisted of two columns of text. The narrative describes each subtest, the appropriate scaled score, and a descriptive classification of the child's performance (for example: low, high, good, fair). Each subtest is described as if it had measured a separate, unique skill as opposed to being thought of as a part of a total battery. No attempt is made to describe any of the scores as being significantly higher or lower than the mean of the test. Instead each receives its descriptive category based on its absolute value in a scale of 1 to 19. (The "Handbook" does provide a description of how one might go about determining subtest strengths and weaknesses, but the way this is done is different than most other programs: Determine the mean of the separate scales, round the results to a whole number, and determine a S/W based on subtests being ±2 points away from the mean). Next a description of IQ and Index Scores is provided. Each score is reported as an obtained score with no confidence intervals noted and no ranges reported. If there is a 15 point difference between the scales, the program provides descriptions of possible reasons for such differences. No mention of base rate is made. Index scores are next reported, whether they are relevant or not and whether or not the scores that make them up differ significantly. The Freedom from Distractibility and Processing Speed Indexes are given solitary descriptions ("ability to concentrate" and "ability to complete timed activity involving the use of nonverbal information"). If an achievement test has been administered, the program compares

both the standard score and grade equivalent to the Verbal, Performance and Full Scale IQ standard score and "expected grade level based on" the respective IQ. (The Mathematics achievement score is alternately called an "arithmetic" and "mathematics" score in the narrative, even if it is Math Computation, Math Reasoning, or a Math Broad cluster score). Finally, the narrative report provides descriptions of up to 33 "Significant Factors." A caution is given about the validity of these factors and if more than 4 factors are 'evident', the program provides a warning to check the scoring of the WISC-III and the entry of data into the program. At the end of the professional report, the program prints a set of "Remedial Recommendations" based solely on the age of the child tested and a perceived weakness (low score) on a particular subtest. Interpretive descriptions, recommended procedures, and suggested materials from various vendors are printed for each suspected weakness.

Changes to the data can be made up to the point of selecting "Complete" from the menu choices. Special procedures are necessary to modify data from a completed report. No text editing is available in the program itself. Files can be saved for editing by a word processing program. There are a limited number of configuration options. The program lacks any ability to set and/or report confidence ranges, to set or report significance levels, and to set or report subtest strengths or weaknesses. Some error checking is available, but it is limited and confusing. If you have inadvertently entered a scaled score that is beyond the range of 19, the program, when you choose "Complete", displays a warning "Too many scores have been skipped" or "Too many scores are out of range". The program does not indicate which score or scores are out or range. The user must review each individual score for correctness.

Although the program was easy to use, its practical relevance is questioned. The interpretive rationale for this program seems to lack much validity. The program's focus on interpreting the exact score obtained instead of the score in relationship to all scores obtained seemed a bit dated. The lack of base rate information, confidence levels (90 or 95%); the lack of confidence bands on global scores; the use of Mental Age and Grade Equivalents all raise concerns about the use of such a program by school psychologists.

(Top of page)

## Report Writer for the WJ-R, Riverside Publishing Company, $348, IBM and Macintosh

The Report Writer for the WJ-R was easily installed onto the Macintosh computers. Entering data (raw scores) was simple and very similar to that of the WJ-COMPUSCORE that anyone using the WJ-R is probably familiar. Moving from screen to screen by clicking on large icon buttons, the user enters various data or chooses various options: background information, raw scores for any of the 21 cognitive subtest and 14 achievement subtest (both forms A and B are available), norms based on age or grade, aptitude/ability achievement discrepancies (choices ranged from 1.3 to 2.3 SD (SEE) Discrepancy), report options of Standard, Customized, or Summary and Table of Scores. With every possible option chosen, the program created a 17 page report that included extensive narrative and a number of pages of score tables. Standard error checking is provided as the user enters raw scores. On screen descriptions provide subtest name and raw score range.

The report is capable of providing detailed descriptions of each test, cluster, and subtest given. Scores are reported in the narrative as grade equivalents, percentile ranks, standard scores, and classification descriptions. In the narrative section of the report, scores are reported as obtained scores with no confidence ranges added. Within the printout for "Table of Scores", the standard scores are reported with 68% confidence bands. There is no option within the program to change these levels to 90 or 95% confidence bands. If cluster scores differ significantly from one another, or if the subtests that make up a cluster differ significantly, the program notes the differences, but gives no plausible explanation for these differences. This was true throughout the program. No real "interpretation" of the scores is made, only the

placing of the appropriate scores within the suitable narrative is done. It appeared to these reviewers that the emphasis on reporting grade- and age-equivalents over the more 'appropriate' standard scores, as well as the limitation of 68% confidence intervals, is unfortunate

The program does provide on screen editing once the program has generated a report. No spell checking capability is incorporated although it was simple to copy and paste the completed report into a word processing program and spell check it from there. (There were no spelling errors found during this review with the exception of some questionable use of words: The report produced the following sentence "Word attack measures Ron's ability in applying **phonic** and structural analysis skills to the pronunciation of phonically regular nonsense words." [BOLD ADDED].

One draw back to the report writer is the lack of graphs that show error bands for each cluster and each subtest that makes up a cluster (similar to those provided with the PROFILES program of COMPUSCORE). These graphs would be useful in determining relevant differences between clusters and within clusters.

Once a report is customized and displayed or printed, you may not go back and re-customize choices without reloading the appropriate case and starting over again. This is not so much a problem as a time consuming process. There is no capability to save as a report as a text file.

This program works only with the Woodcock tests. If the Woodcock Cognitive is given, the only achievement scores and comparison available is to the Woodcock achievement.

This program does a wonderful job for the tasks it is designed to do. The biggest limitation of the program is the interpretations given. The program does a good job developing narrative reports but the reports lack any true interpretation. No effort to describe meaningful differences between clusters or subtest s are made and no real hypothesis generation is made.

(Top of page)

## Bottom Line

Although all the reports are editable in one way or another and do, to differing degrees, contain error checking, these authors did find significant differences between the programs' rationale and output. Each program, in its own right, offers comparable yet differing levels of interpretation; completeness of reports; and number of tests available for interpretation. Practitioners will want to assess their own individual needs when choosing among these programs. For example, only the WISC-III Writer (TCP) and the Report Writer for the WJ-R offers a program that can convert the raw scores to scaled scores for the appropriate tests, but these programs are limited to the WISC-III and WIAT and WJ-R Cognitive and Achievement respectively. The Ewing and Psychological Support Systems programs offer the most comprehensive number of tests allowable, but often do so at the expense of real interpretive and accurate statements. For devotees of Alan Kaufman's method of interpretation, his KWIIS is a good choice, but it too is limited. The most accurate test results were found in programs published by the companies that also publish the respective tests, but they are typically limited to those single tests. (It should also be noted that these reviewers discovered a mistake in the 'Shared Abilities' report created by the KAIT-ASSIST program published by AGS.) For a program that offers multiple tests, the Ewing program provides a large number of option in an attractive, easy to use program. The EAW3 is very different from most of the programs reviewed, both in the interpretive methods used and the format of the report. Users of this program should understand and agree with the Nicholson and Alcorn analysis that is reported. (The first author has major concerns about the accuracy and

validity of this method).

No matter what program a school psychologist chooses, the user is ultimately responsible, both legally and ethically, for the reports generated. Each program has, either in the manual or in the program itself, a disclaimer to the user about this point.

These reviews, comments and opinions are those of the authors and may not reflect that of NASP, the editorial board of the Communiqué, and/or the affiliations of the authors.

(Top of page)

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792391 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Dumont, R., & Chafouleas, S. (1999) Conducting behavioral observations: Some technical support?, Communiqué, Vol. 27, 7, 32-33

Since the proposed IDEA regulations now require careful analysis of behavior in certain prescribed situations, school psychologists are looking for more efficient ways to conduct and collect data from behavioral observations. The use of computers and particularly specific software may aid in the collection and analysis of this data. Two recently published software programs, !Observe and the Behavior Observation Assistant (BOA), purport to make the collection, organization, and storage of behavioral observations trouble-free. These programs have been used and critiqued. Information specific to each program is provided, as well as summary comments regarding the use of computer software for data collection.

**!Observe software ($149), Psychsoft, Inc. (1-800-536-4996)** http//www.psycsoft.com

!Observe software is available for all Windows and Mac platforms. The software is also available for the Apple Newton (110 or better). The !Observe computer program (version 1.0b4) was reviewed using a Macintosh PowerMac 7100/66, a Power Book 160, a Newton Message Pad 130, and an IBM Dell portable. Installation on each computer was straightforward and flawless. Simply copying the appropriate files to the Mac hard drive or running the Setup program on the IBM had the files and programs installed in seconds. During this review, we found that use of the Apple Newton (or any similar IBM based palm size computer) had a number of distinct advantages over the laptops and desktop computers. Due to its small size and unobtrusiveness, the Newton was ideal for taking to a classroom to record behaviors. Since familiarity with the mouse, trackball, or touchpad is necessary for input of the data, the Newton, with its stylus input, made using the software efficient and easy.

Although a short 20+ page manual accompanies the software, anyone semi-computer literate will find this software very user friendly. The data collection process is built around customizable templates that we found to be extremely easy to produce. Once familiar with the program, in less than 5 minutes, these reviewers were able to create a rather elaborate template for comparing various on-task / off-task behaviors for a subject and a control. Templates, either provided with the software (ED, MR, ADHD, Critical Incidence, Autism, Mental Status, and a Functional Assessment) or created by the user, are simply groups of "behavior buttons" that are used to tabulate which behaviors will be observed. The templates that come with the program were useful as examples of what could be easily done with the !Observe software. We found that it was so easy to create templates that we never used any of the preprogrammed ones. Each template may have up to 24 "behavior buttons which may be color-code buttons to aid collection. Whenever a behavior of interest is noted, the observer simply clicks on the

appropriate button using the stylus, mouse, track pad, or pointer device. Each behavior may be grouped into some broader variable (i.e., Class and Category). For example, a button labeled "hitting" might be coded as belonging to a negative class and a physical category while another button labeled "quiet work" might be coded as belonging in the positive class and the nonverbal category.

Cues are available to serve as a reminder to enter your observation data. Time interval may be set to any whole-second increment , and total observation time can be set in any whole minute increment. Reminders are provided by either a Flash (the background behind the headings of the capture window change color every time an interval passes) or a Beep (the system beep will sound when an interval passes). Users may choose to have both or neither reminders on. Although the author lists a number of data collection options, we found the program capable of being configured to aid in a variety of observation formats. Specifically, frequency counts, interval and duration recording, and momentary time sampling were all readily available. In order to conduct each type of observation, the user simply needs to plan at which point data is to be recorded, and then consider the collection format when interpreting the summary report.

The program automatically saves the observation data. The summary report lists each of the observed, coded behaviors followed by their class, a count of the number of times each of the specific behavior buttons was clicked, and percentage of the time the button was pushed in relation to all other buttons. In addition, it is possible to obtain information regarding rate per minute and duration (in seconds) of each behavior. Users may also elect to have observations summarized as a data stream, listing each behavior in the order of recording along with the exact time (8:43:44 PM) that the behavior occurred. !Observe provides two forms (Standard or Interval) for summarizing the data. Standard form records every behavior button pushed during a set interval while Interval form tracks every different button pushed within an interval. If, for example, within a single 10 second interval the "On Task" behavior button is pushed three times and the "Off Task" behavior button 2 two times, !Observe would record this as On Task - 60%, Off Task - 40% using Standard form and On Task 50% (one interval) and Off-Task 50% (one interval) using interval form. These reviewers saw little utility in the information provided by the Interval summaries. Bar and Pie charts are available.

**Behavior Observation Assistant, ($99.00) Bunger Solutions (972- 424-9647)** http:www.bungersolutions.com

The Behavior Observation Assistant (BOA) is available for IBM or IBM compatible platforms, and requires Windows 3 or later to operate. The software comes on 5 3 ½ inch disks, and is easily installed in minutes. Technical support is available through an online help section or via email. The accompanying manual is easy to read and provides a number of useful examples. Although the largest feature of the program is data collection and summary, the Behavior Observation Assistant also provides a template for creating behavior intervention plans, and reference tools related to behavior management (i.e., manifestation determination, checklists of possible reinforcers).

The setup for data collection is easily accomplished through the use of an on-screen toolbar. Data collection itself involves the following steps: subject setup, selection of behavior(s), and collection of the data. Subject setup is done by first selecting a group (i.e., Mr. Smith's classroom) and then an individual (Johnny) within the group. Preprogrammed target behaviors are provided (i.e., hitting, vision not properly directed, speaks without permission). These may be easily selected and edited by clicking on each or the observer may edit or create totally new behaviors. Observation location (i.e., gym) and setting can be specified. Up to 6 behaviors can be recorded when observing 1 subject, however if more than 1 subject is involved in the observation, only 1 behavior may be recorded. To actually collect data, the observer presses preset function keys that relate to the selected behaviors. Pressing the escape button allows the observer to make comments or correct errors. However, doing this halts the observation, thus, providing inconsistencies in observation real time when comparing sessions.

The authors report a number of data collection options, including interval recording, time sampling, frequency recording, and duration recording. The choice of observation type (i.e., baseline, intervention, follow-up) is also provided as an option. Observation time periods and other collection information can be modified according to need, however, range limitations may be present depending on the type of observation conducted. For example, time can be selected from 5 seconds to 1 minute when using interval recording while time periods from 1 to 10 minutes are available when using time sampling. It should be noted that the BOA's time sampling procedures are not equivalent to momentary time sampling procedures described by Saudargas and Lentz (1986). The BOA stops recording time until either a response is entered or 10 seconds have elapsed before moving on, thus potentially providing inconsistency in the collection across sessions. Data from each observation is saved, and may be viewed in a summary report. The summary report presents all background information regarding the session (i.e., Who was observed?, How were they observed?, How long were they observed?), and combines the observation data in table format. Reports may be exported into a file format (i.e., word processor) different from BOA. Observation data is automatically saved and sessions may be organized through the file management button on the toolbar. Graphing options are not available, and data across observation sessions cannot be combined into one report.

Summary Comments

In summary, the !Observe and the Behavior Observation Assistant provide a variety of data collection and behavior management options in a relatively easy to use format. Subject and behavior setup is easy to accomplish, and the behavior choices provided should fit the needs of most school psychologists. In order to become fluent with how to record the data, we found it necessary to practice a number of times the use of the input device. For the BOA, if 6 behaviors have been coded into the template, it may be difficult to remember which function key corresponded to which behavior. With the !Observe software, the buttons themselves are labeled, making the input very smooth. However, we did notice that if we used templates with many labeled buttons, we had some difficulty keeping accurate track of behaviors simply because we had too many buttons. The more we used the two programs, the less of a problem we had. For both programs, the user must carefully consider how data collection is configured in order to understand the type of observation procedure he/she is using, particularly when comparing information across subjects and observations. For example, the !Observe manual states that interval observations are conducted, but the description of the data collection actually suggests momentary time sampling. The BOA describes time sampling, but this procedure is not the same as the momentary time sampling conducted by the !Observe. Bell and Beedle (1993) provide a review of data collection techniques and terminology. For both programs, the observation summary reports are easy to read, and save time by doing all tabulations for the observer, but multiple observations need to be combined by hand. Finally, for the BOA, although it can be helpful to have references related to behavior management and a behavior intervention plan template, the observation data are not linked to the template. For both programs, technical assistance, by phone and/or e-mail, was quick, professional, and accurate.

Is it worth using computer software programs to make behavioral observations? The answer is - it depends. First, use of the software in school settings is may be limited by the technical requirements. Clearly, one needs at least a laptop computer to fully utilize either software program (we can't imagine anyone hauling a desktop computer into a classroom). Until students were familiar with our presence, carrying a laptop computer into a classroom invariably caused a stir and a number of raised eyebrows. Nonetheless, the laptop served its purpose and the observations went smoothly. Second, usefulness depends on the purpose of data collection and desired efficiency. If the purpose is to collect frequency counts of a behavior, it may not be more efficient to bring out a computer when paper would do just as well. For more complex data collection techniques such as momentary time sampling, use of the software by an observer may facilitate accuracy of the collection given the potential auditory and visual reminders, and save time with data tabulation. However, since the observation reports do not easily combine data across observations, the observer does not save time with the analysis of behavior patterns. In summary, we were able to identify a number of positive features of each of the reviewed software programs which may be useful in the school setting. Software

selection is best made after careful analysis of the purposes for observation.

References

Saudargas, R. A. & Lentz, F. E. (1986). Estimating percent of time and rate via direct observation: A suggested observational procedure and format. School Psychology Review, 15, 36-48.

Bell, D. R. & Beedle, B. B. (1993). Observing and Recording Children's Behavior. Kendall/Hunt Publishing

Ron Dumont is Associate Professor of Psychology and Director of the MA and Psy.D. programs in School Psychology at Fairleigh Dickinson University. Ron is also a contributing editor for the Communiqué

Sandy Chafouleas is Assistant Professors of Psychology and director of the SUNY-Plattsburgh School Psychology program..

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792392
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Learning Efficiency Test-II (1992 revision)

This is not meant to be a comprehensive review of the LET-II but merely a summation of questions and concern.

The Learning Efficiency Test-II (1992 revision) is published by Academic Therapy Publications and authored by Raymond Wheeler, a Professor of Psychology at East Carolina University in Greenville, North Carolina. It costs $60.00 for a manual (187 pages), stimulus cards, 50 record forms, and a vinyl folder. Norms are provided for ages 5 through 75 and older, with tables presented in one year intervals between 5 and 16 years.

The catalog and manual claim that it is a quick and reliable measure of visual and auditory memory characteristics and that it can provide useful information about a person's preferred modality of learning, as well as providing information about the impact of interference on memory storage and retrieval.

The LET-II can be administered in 10-15 minutes. It consists of strings of 2 to 9 non-rhyming letters presented either orally or visually. The child/adult responds verbally and subtests raw scores are based on string length. The raw scores are then converted to scaled scores and percentile ranks. Memory on the LET-II is assessed in two modes (visual and auditory) and in three recall conditions (Immediate, short-term, and long-term). The six subtest scores (Mean 10, SD 3) can be converted into Modality scores as well as a Global Memory score (Mean 100 SD 15).

Given the way that memory can affect a child's functioning in school, and the need to assess preferred learning modality, the LET-II is a test that we might be tempted to buy.

An examination of the LET-II tests material, manual, and catalog raised the following concerns:

The manual states that the LET-II "has been empirically demonstrated to be highly predictive of actual classroom levels of performance in reading and mathematics for students with average ability as well as handicapped students" (page 10). The support for this claim is 4 citations. Examining the bibliography at the back of the manual, these four empirical studies turn out to be 3 unpublished Master's degree theses done at East Carolina University in Greenville, North Carolina. The fourth citation is for the LET (1981) manual.

Standardization:

The LET-II was normed on a sample of 1126 children and adults between the ages of 5 years 0 months and 85 years, 4 months. Data is provided for the age, sex and race of the sample but no Socio-economic status data is provided. The manual states that the "participants came from a broad range" of SES backgrounds. Sex variables are fairly even across the sample with 46% male and 53% females being included. This is very close to the U.S. Census Bureau's 1990 estimates of 48% and 52% respectively. Race variables are less comparable to the Census

data. The LET-II reports percentages of 66 versus 33 for Caucasians and Blacks in the total sample. This in contrast to the 77 and 12 percent estimates for the U.S. population. For the ages 5 to 16 the percentages are 58 versus 42 for Caucasians and Blacks. No geographical data is provided. It is my assumption that most came from North Carolina.

For each of the one year age intervals, 5 through 16 (not 5 through 15 as stated on page 7 of the manual), the samples average only 55 people, with a high of 84 at age 8 and a low of 40 at age 16. These are well below the sample size of 100 recommended by Salvia and Yssledyke (Assessment in Special and Remedial Education/Third Edition, 1985) for the computation of standard scores.

No data about the IQs of the children in the normative sample is provided. Adults in the sample were administered the Peabody Picture Vocabulary Test and anyone scoring below 85 were excluded from the sample. It is further noted that "no known cases of mental retardation as defined by a general IQ score of less than 85 (on either a group or individually administered intelligence test) were included in the sample." (italics added). Two points. Was this the PPVT or the PPVT-R and would the age of the norms have an effect on the IQ? Who defined mental retardation as "a general IQ score of less than 85"?

Reliability:

Reliability studies using the LET-II were not conducted. Information based on 2 studies using the LET (1981) are included in the manual as measures of the LET-II reliability. One unpublished, "informal" study, involved 55 learning disabled student in grades 4 through 12. Coefficients for this group were found to be .71 to .86 (median .80). No breakdown by age, grade, IQ, or time interval is provided in the manual. Without that information, and the interval between test and retest, this data seems almost meaningless. A second study was done which involved only 40 students identified as having "learning and behavior problems." These students each had IQs above 89 and were test-retested between 1 to 6 weeks. Reliability coefficients ranged from .81 to .97 and are generally considered adequate. However, there are no specific information about age and or grade of the subjects. All are simply labeled "secondary" students in the manual. Since these reliability studies were carried out using the LET and not the LET-II, no coefficients, and hence no Standard Error of Measurements are provided for the two Modality scores or for the Global Memory score. Without this information, there is no way to calculate meaningful differences between the Modality and Global memory scores. The reported coefficients also offer no information from which to draw conclusions about the permanency of the scores since the time limit between the two administrations of the test was fairly short.

Validity:

Validity in the manual is addressed in a number of ways: content validity, diagnostic validity, and predictive validity.

Diagnostic validity was shown by examining patterns of performance among 4 groups of special education students and comparing these patterns to a group of "average" students in grades 4 though 7. The 197 Average students had average Verbal, Performance, and Full Scale WISC-R IQs of 111, 110, and 112 respectively. This raises some question about what might be expected when this 'average' group is compared to a special education population. Average is typically thought of as having a mean of 100, not 112! Won't comparing a group labeled "LD" who have a mean IQ of 93 with a group labeled "average" with a mean IQ of 111 create some confusion when interpreting? My average and this sample's average are different.

One confusing aspect of the statistical tables presented in the manual is the arrangement of the categories for the "learning groups." In the

tables listing Correlations, Mean recall characteristics, Percent of information loss, Reading and Math achievement, and Stepwise and Simultaneous regression, the order in which the categories are presented change from page to page. The first table lists the order as Average, LD, EH, EMH and SL; yet the very next table lists the order as Average, EH, LD, EMH, and SL. In the next 3 tables, the order changes again and the EMH group becomes an EMR group? One may quibble that this is a minor point, but in a manual that contains other errors, some fairly serious, this does raise questions about the validity of such tables and statistics.

Table 2, which presents 4 pages of Means and Standard Deviations for raw scores based on string length by age level grouping, contains either a very curious developmental anomaly or a very serious error. Page 163 is a table for Auditory Memory-Ordered. For the Immediate and Long term trials the table shows the expected increasing lengths of the string as the person's age increases. A child of 5 remembers an average 2.91 letters while a person of 16 remembers 5.41. However, for the column for the Short term recall trial we discover that all ages had mean recalls of approximately 1 letter?? In fact, according to this chart, at age 70-74, a person remembers 1.71 letters but the standard deviation is larger than the mean (2.04)? This is almost certainly a serious error in the table. How many of the other tables are in error? I don't know for sure, but curiously, looking at the conversion tables for raw score to scaled score for a 15 year old (page 99) I found another! As you examine the table you note that the higher the raw score one has, typically the higher the scaled score will be. Not so for age 15, unordered, immediate recall. OOPS!

Table 6 (page 167) and Table 7 (page 168) display mean recall characteristics and percent of information lost for each recall trial for each special group. These pages are interesting for the way the information is explained in the text of the manual. The manual explains that special groups show distinct patterns of differences in recall capacity. The tables do show this but one must ask about the use of percentages as the measure of loss. For example, the average child, with an IQ of 112, starts by remembering 4.7 letters and drops to 3.7 on short term recall, a loss of 22.3%. The LD child starts at 3.5 letters and drops to 2.2 on short term recall. A drop of 37%. The percentages are different by such large percentages only because the LD child started at a lower number. The actual loss is about the same, 1 letter!

One aspect of validity not addressed in the LET-II manual is that of carry over to real life. The LET-II has items that are discrete units of abstract symbols (letter) delivered in a non meaningful manner. With regard to children in school, much of the material to be learned requires the recognition and use of meaningful items. Krupski (1985) demonstrated that for learning disabled children, memory performance is approximately equal to that of average students if the material being presented is meaningful in nature. If this is true, the LET-II may be measuring some important aspect of non meaningful memory, but tells us nothing about the child's performance in real life situations.

Since the LET-II proposes to measure and relate certain learning styles to academic functioning, there should be some studies to show the meaningful relationship of different learning styles to classroom success. The manual does not offer any study to support any claim that academic deficiencies can be attributed to a weaker modality score on the LET-II.

Administration and Scoring:

Administration of the LET-II should be fairly straight forward and trouble free. Unfortunately, it isn't. Although the manual presents detailed instruction on how to administer and score the LET-II, these instructions are at times unclear and worse, they contain errors!

When administering the LET-II immediate recall subtests, the examiner presents the visual stimulus items for 2 seconds while the Auditory stimulus items are read 1 per second. After the child repeats the items back to the examiner, s/he is asked to perform some interference task

("Count from 2 to 12"). The manual states that some children have difficulty with the counting interference task, but examiners are not instructed what to do if the person can't perform the task.

Recording responses is addressed on pages 36-37 of the manual. Unfortunately the examples used on these pages contain errors in the scoring. (In fact, there are two errors in this section). When explaining how to score an item for which the child has extended the string beyond the length, the examiner is told to disregard any letter beyond the proper length. The example given is for a child presented the string "Q-R-H-X"and who responds "Q-R-X-H-Y." Scoring is for both ordered (correct placement of letter) and unordered (correct letter, regardless of placement). The manual states: "Scoring of items would end with the fourth item in the student's response. Both Ordered and Unordered raw score would be four." (italics added). Gee, I thought the score would be 2 and 4 respectively?

The second error on the page is in the notation of the sequence. Letters that a child gives that are not in the presented string are circled to indicate such an error has occurred. The examples on page 37 have letters that were included in the string circled to indicate that they were not included in the string!! But wait...the same notation error is made again on page 72 for a different protocol! And on page 76, another error, but this time in a different notation. Can anyone get this right?

Okay, I'm done with errors, but I stopped looking.

How about just wanting to be average. Examining the norms tables for ages 5 to 20, it is interesting to note that a child cannot obtain an "average" scaled score of 10 a large percentage of the time. For example, in the Visual modality tables, there is no raw score to scaled score equivalent of 10 in 58% of the cases. In the Auditory Modality tables the child will fare better. Here there is no raw score to scaled score equivalent of 10 in 56% of the cases. What do raw scores convert to then if not average? Often the converted score is 8 or 11. A 3 point difference based on 1 raw score point. Three points! One standard deviation!

So what's my point? I guess it's that you can't tell a test by it's advertising. No matter how good a test looks; or what company publishes it; or that it is published at all, we as examiners are often left to critique the tests we buy. In fact, that is our ethical responsibility. If we use it , we are responsible for defending the reliability and validity of it's use. After spending lots of money on a test, it's too bad how often we can be disappointed by the product.

As far as the LET-II is concerned, I'd follow Lennon and McCartney's refrain "Let it be. Let it be." As for all others, Caveat Emptor (Buyer beware).

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792393 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Kaufman Adolescent and Adult Intelligence Test

Kaufman, A. S. & Kaufman, N. L. (1992). Circle Pines, MN: American Guidance Service

For a complete review, see:

Dumont, R. & Hagberg, C. (1994) Kaufman Adolescent and Adult Intelligence Test (KAIT): Test Review *Journal of Psychoeducational Assessment*, 12, 2, 190-196

Alan and Nadeen Kaufman, through American Guidance Service (AGS), have released the "new kid on the block" in intelligence testing: the Kaufman Adolescent and Adult Intelligence Test (KAIT) ($495). This individually administered intelligence test is designed to assess persons 11 to 85+ and is composed of separate Crystallized and Fluid scales. The theoretical model for the scale was derived from Horn and Cattell's (1966, 1967) conceptualization of Fluid vs. Crystallized Intelligence while models for the 8 subtests included Luria and Golden's (1981, 1980) planning ability and Piaget's (1972) formal operations. Three subtests (Definitions, Auditory Comprehension, and Double Meanings) form the Crystallized scale while three more (Rebus Learning, Logical Steps, and Mystery Code) form the Fluid scale. An additional subtest for each scale (Famous Faces for the Crystallized and Memory for Block Designs for Fluid) are available as a substitute to one of the core subtests. Two measures of delayed recall (Rebus recall and Auditory Comprehension recall) permit the comparison of performance across time. Also included with the KAIT is a normed 10-item Mental Status subtest.

The test material comes in an attractive, locking briefcase which contains 2 easels; wooden blocks for the Memory for Block Designs subtest; an audio cassette for the Auditory Comprehension subtest (the user supplies tape deck); protocols for the test including separate Mystery Code booklets; and the test manual.

## Standardization

The KAIT was standardized on a sample of 2000 people selected as representative of the US population on the basis of the 1990 population estimates. 13 age groups, ranging from 11 to 75+, are divided by geographic area, race and ethnic group, and examinee or parental education levels. The size of the sample varies by age. For the children's sample, age 11 through 16, 500 children were tested, for the adult sample, age 17 through 75+, 1500 were tested.

## Deviation IQs and Scaled Scores

The IQ tables in the KAIT manual are based on only 6 of the 8 subtests. The two alternative subtests, Famous Faces and Memory for Block Designs, are excluded from the calculation of the IQ except if they replace a subtest not administered. Prorating of either scale IQs or Composite IQ is discouraged.

## Reliability

Reliabilities for the KAIT are generally outstanding. Each of the 3 IQ scales has an internal consistency reliability that averages .95. The six core subtests have an average reliability of .90. Test-retest reliability for the Crystallized, Fluid, and Composite IQs were .94, .87, and .94, respectively. The subtest reliabilities are not as robust. The average test retest reliability for 5 of the eight subtests was found to be in the .70s. Because of these 'low' reliabilities, interpretation of the KAIT should focus on subtest groupings and not on individual subtests.

## Validity

Construct validity for the KAIT's Fluid-Crystallized make-up was examined in a number of ways. Exploratory factor analysis found support for only these two meaningful factors. Each subtest conformed to its' hypothesized Crystallized vs. Fluid assignment.

Construct validity was also determined by comparing the scores obtained on the KAIT with those obtained on various other measures of intelligence, including the WISC-R (Wechsler, 1974), WAIS-R (Wechsler, 1981), SB-IV (Thorndike, Hagen, & Sattler, 1986), and the KABC (Kaufman & Kaufman, 1983). In general the KAIT was found to have high correlations, and thus, substantial variance overlap with the Wechsler tests and the Stanford Binet, while having lower correlations with the KABC mental processing composite.

Diagnostic validity studies were carried out with various clinical samples. These included Neurologically Impaired, Clinically Depressed, Alzheimer's Type Dementia, and Reading disabled. For each clinical sample a Control group was created from the standardization group by matching the clinical samples' age, gender, race, and years of education. These studies offer the examiner and researcher interesting interpretive hypotheses.

There are no comparisons for either a mentally retarded or gifted population. This might have been especially useful for the MR group since the studies by Zimmerman, Covin, and Woo-Sam (1986) and Rubin, Goldman, and Rosenfeld (1985) found the IQ classifications based on the administration of a WAIS-R to frequently change from those classifications found from the administration of a WISC-R. Another sample not included that might have been helpful was a learning disabled group. The manual did include a Reading Disabled sample, but this included only 14 children. A sample for the construct validity study also included a Learning Disabled sample, but this group was comprised of only 8 children. School district personnel using the KAIT will have to wait for research to determine its usefulness in identification of learning disabilities. One other concern with the clinical sampling was the inclusion of such a small number of children in the neurologically impaired group without clearly delineating the age breakdown in the manual. The total group contained 44 people. Of these, only 3 (two 11-year-olds, and one 17-year-old) were below the age of 20 (Colin Elliott, personal communication, March 1993). This may lead some to over interpret the results. The KAIT may be found to be very useful in differentiating various problems in neuropsychological, and psychoeducational samples. Clearly, more research is needed and expected.

## Administration and Scoring

Administration time for the Core Battery (6 subtests) is approximately one hour, with the Expanded battery (6 core subtests, 2 alternative subtests, and 2 recall subtests) taking approximately 90 minutes. Directions for each subtest are contained on the easel pages, so no additional manuals are needed during administration. Subtests are given in the order presented in the easels.

All subtests, with the exception of the delayed recall subtests and Famous Faces, provide teaching tasks to ensure that the examinee fully understands the nature of what is expected. Teaching items allows the examiner flexibility in choosing alternative wording or method to ensure that the person tested understands each task.

Although a stop watch is necessary during the administration of the KAIT, no items are given 'bonus' points for successful completion within time limits. Several subtests require a response within a certain amount of time, and the designs for the Memory for Block Designs are shown for only 5 seconds.

Everyone administered the KAIT starts with the teaching items and begins formal testing with item number 1. No different age differentiated starting points are used. Discontinuation rules vary from subtest to subtest, and are clearly noted on the individual record form.

Because certain subtests require skills (reading, adequate hearing, etc.) that may prevent a person from adequately demonstrating ability on the KAIT, the easels provide cautions to examiners about which subtests might be omitted from the core battery because of such problems. The decision to omit a subtest needs to be made carefully and with good reason.

Administration flows smoothly from subtest to subtest. Recording responses on the protocol is simple and straight forward. Most verbal items require a simple word or number response. Scoring items as they are presented is necessary for discontinuation decisions.

Interpretation

Since the use of the KAIT is recommended for those legally and professionally competent to give existing intellectual assessments, some level of interpretive skill is expected of examiners. Four chapters of the manual provide background information about the subtests and theoretical information to aid interpretation.

The protocol provides tables used to determine, by age categories, the significance of the difference found between the Crystallized and Fluid scales. If the separate scales do differ significantly, base rate information is provided in the manual to determine frequency and unusualness of such scatter.

The last page of the protocol provides tables for comparing each subtest to the mean of the relevant scale. Three tables in the manual provide the actual difference needed for significance carried to 1 decimal point.

Because of the low test-retest reliabilities of 5 of the 8 subtests, examiners are instructed to focus interpretations on subtest groupings as opposed to individual subtests. (Each subtest does contain adequate levels of subtest specificity to be interpretable, but groupings are thought to be more stable.) Tables in the manual provide hypothesized groupings for both "Shared abilities" and "Influences affecting Performance."

Conclusion

The KAIT appears to offer an exciting new addition in psychologists' choice of intellectual assessment tools. Given the KAIT's age range (11 to 85+), it can easily substitute for a WISC-III at the adolescent level, and the WAIS-R at the adult level.

In the process of reviewing the test, the reviewer administered over thirty KAIT's. Test procedures were easily learned. The KAIT manual was very helpful. Both children and adults found the test to be non threatening, and for some, enjoyable and challenging.

Overall the test seems to be well thought-out and validated. This new test will probably become a test of choice for some researchers, especially those interested in assessing the geriatric population. The Kaufman's have gone to great lengths to standardize a reliable, and valid measure for older Americans. Valuable research should be conducted using the KAIT with school aged populations so that its merit for school psychologists can be determined.


Golden, C. J. (1981). The Luria-Nebraska Children's Battery: Theory and Formulation. In G. W. Hynd & J. E. Obrzut (Eds.), Neuropsychological assessment and the school-age child: issues and procedures (pp. 277-302). New York: Grune & Stratton.

Hansen, J. C., & Campbell, D. P. (1985). Manual for the SVIB-SCII (4th ed.). Stanford, CA: Stanford University Press.

Horn, J. L., & Cattell, R.B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. Journal of Educational Psychology, 57, 253-270.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. Acta Psychologia, 26, 107-129.

Kaufman, A. S., & Kaufman, N. L. (1983). K-ABC Interpretive Manual. Circle Pines, MN: American Guidance Service.

Luria, A. R., (1980). Higher cortical functions in man (2nd ed.). New York: Basic Books.

Piaget, J. (1972). Intellectual evolution from adolescence to adulthood. Human Development, 15, 1-12.

Rubin, H. H., Goldman, J. J., & Rosenfeld, J. G. (1985). A comparison of WISC-R and WAIS-R IQs in a mentally retarded residential population. Psychology in the Schools, 22, 392-397.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). Stanford-Binet Intelligence Scale: Fourth Edition. Chicago: Riverside.

Wechsler, D. (1974). Manual for the Wechsler Intelligence Scale for Children-Revised. San Antonio: The Psychological Corporation.

Wechsler, D. (1981). Manual for the Wechsler Adult Intelligence Scale-Revised. San Antonio: The Psychological Corporation.

Zimmerman, I. L., Covin, T. M., & Woo-Sam, J. M. (1986). A longitudinal comparison of the WISC-R and WAIS-R. Psychology in the Schools, 23, 148-151.



**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792394
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Test of Memory and Learning

## Reynolds, C.R. & Bigler, E.D. (1994). Austin, Tx: Pro-Ed.

For a comprehensive review see:

Dumont, R., Whelley, P., Comtois, R., & Levine, B. (1994) Test of Memory and Learning (TOMAL): Test Review *Journal of Psychoeducational Assessment*, 12, 2, 414-423

Drs. Cecil Reynolds and Erin Bigler, through Pro-Ed, have released the Test of Memory and Learning (TOMAL) ($159). It is designed to assess persons 5 through 19 years of age and is composed of Verbal , Nonverbal , and Composite Memory Indexes. An Delayed Recall Index may be computed from the scores of 4 recall subtests. Descriptions of terms used to define memory are adapted from the work of Kolb and Whishaw (1990) and L. Squire (1987). Five subtests form the Verbal Memory Index while five more form the Nonverbal Memory Index. Additional subtests for each Index are available as additions to the core subtests. Four measures of delayed recall, given 30 minutes after the start of testing, permit the comparison of performance across time. Also included with the TOMAL are directions for computing 5 Supplementary Indexes (Attention/Concentration, Sequential Recall, Free Recall, Associative Recall, and Learning). Finally, normative tables are provided on the Supplementary Analysis forms so that a person's learning and retention curves can be drawn and compared.

Test material comes in a box which contains: 2 booklets containing the stimuli for subtests; Facial Memory chips; Visual Selective Reminding test board; Delayed Recall cue card set; protocols for test including separate Supplementary Analysis forms; and the examiner's manual. The booklets do not stand on their own, and an easel is recommended by the publisher. The picture stimuli books were cumbersome when flipping from page to page. Tabs would have been helpful for selecting the correct pages to turn to. Because some of the pages are printed on both sides, laying the booklet flat on the table allowed the child to see, and often be distracted by the material on those pages.

Standardization

The TOMAL was standardized on 1342 people selected as representative of the US, 1990 population estimates. 15 different age groups were divided by geographic area (17 states), race, gender, and ethnic group. Socioeconomic Status (SES) was determined on the basis of the test site chosen. The sample was also stratified by urban versus rural residence. The size of the 15 sample ages tested ranged from a low of 42 at age 18 to a high of 163 at age 11.

Reliability

Internal reliabilities for the TOMAL are generally high, ranging from a low of .56 to a high of .98. Of the 4 core Indexes, all but Delayed

Recall have reliabilities in the 90's. Delayed Recall averages .85 across all ages. Each of the 5 Supplementary Indexes have reliability estimates in the .90's. Nine of the fourteen core subtests have an average reliability in the .90's while the remaining 5 are in the 80's. Delayed recall subtests generally had lower reliabilities than did the other subtests.

Test-retest reliability coefficients for the Core and Supplementary Indexes ranged from .81 to .92. The fourteen subtest reliabilities averaged .81. Average test-retest reliability for 3 of the fourteen subtests was found to be in the .70s. Two of these subtests, Facial Memory and Abstract Visual Memory, are primary subtests of the Nonverbal Memory Composite.

Validity

Content validity was assessed by reference to initial development and tryouts, task analysis, and representativeness. Because of the newness of the TOMAL and the lack of agreement as to the definition of the construct, the authors state "Until the field can coalesce the innumerable designations of forms and types of memory, an empirically verifiable analysis of content validity under this definition will always go wanting."

The exploratory factor analysis found that the TOMAL subtests all had positive correlations with every other TOMAL subtest. The results of factor analysis show strong evidence for a measure of general memory and thus for the TOMAL's Composite Memory Index (CMI). A Two Factor solution did not support the Verbal/Nonverbal index interpretation. The Four Factor solution did not lend support for the Verbal/Nonverbal division of the TOMAL. The process determined groupings are said to supplement the "content-driven" and "expert derived" indexes.

The TOMAL showed low correlations to measures of achievement, typically about .10 points below what is typical of IQ/Achievement comparisons. When compared to the WISC-R and the K-ABC, the TOMAL correlated in the mid .50's. No comparisons between the TOMAL and other measures of memory, such as the Wide Range Assessment of Memory and Learning (WRAML, 1990) are included in the manual.

A single diagnostic validity study was reported for a sample of learning disabled children.

Administration and Scoring

Administration time for the Core Battery (10 subtests) is approximately 45 minutes while the entire battery (10 core subtests, 4 alternative subtests, and 4 recall subtests) takes approximately 60 to 75 minutes. Directions for each subtest are contained on the protocol pages, so the manual is not needed during administration. The Facial Memory, Paired Recall, Digits Backwards, and Letters Backwards subtests are the only subtests to provide unscored teaching tasks. Teaching items are also allowable on a number of other subtests, but only after the person has attempted an item and received a score of 0.

Seven of the 10 core subtests have different starting or stopping points depending on the age of the person. Discontinuation rules vary from subtest to subtest, and are noted on the individual record form.

Administration flows smoothly from subtest to subtest. Recording responses on the protocol is straight forward but certainly not without some difficulty. Without considerable practice, examiners will find this task quite difficult on a number of subtests.

A number of specific questions that seem unanswered by the manual were found during these authors' administrations of the TOMAL. The manual states that for the Memory for Stories (MFS) subtest, a child is given "1 point for each element of the story repeated correctly." Some children tested simply repeated back to the examiner correct elements, but missequenced them, resulting in correct recall of facts but a total misunderstanding of the story content. Does a child who recalls only salient facts at the expense of the story comprehension receive credit? Must the answers be in some logical sequential manner? Along those same lines, there may be confusion on the subjects' part about how they are expected to retell the story. For the first story the child is instructed "I'm going to tell you a story. Listen carefully, because when the story is done, I want you to tell me everything you remember about the story." The story is then read to them. Only after this reading is the child instructed to "Tell the story back to me the very best you can". The need for verbatim responses is not clearly conveyed in the initial instructions. It is only later in the second and third stories that the child is instructed "Tell it back to me just the way you heard it." One subject asked "Do you want me to use your words?" This suggests that the manual is not clear enough, at least for the persons taking the test. There is no direction for the examiner to query possibly ambiguous responses. There are guide words listed in the table but no criteria is given as to how to judge accuracy of close but not verbatim words, colloquialisms, and synonyms.

Interpretation

Since the use of the TOMAL is recommended for those considered professionally competent to give assessments, some level of interpretive skill is expected of examiners. Three chapters of the manual provides background information about the subtests and theoretical information to aid in interpretation of the test. Tables in the manual provide information about the frequency of differences when certain comparisons are made. Such baseline data are helpful in determining the relative frequencies of differences and thus making logical statements about the abnormality of a suspected difference.

Regarding the meaning of scaled scores, the TOMAL authors state: "Because a scaled score of 10 is average on the TOMAL, scaled scores higher than this average indicate a strength, and scaled scores below this average indicate a weakness, or impairment, relative to age peers." (p. 41). For years, psychologists have been taught that reliance on scaled scores are suspect since they are unstable. All scores contain error, and because of that a scaled score of 9, although "below the mean" and supposedly indicating "impairment" on the TOMAL, must be viewed with caution. The inclusion of this sort of interpretive guideline in the manual for the TOMAL may lead some to over-interpret scores that are otherwise meaningless deviations from the mean.

A second point of apprehension these reviewers had is regarding Global score comparisons. When discussing the Global Scale Comparisons, the TOMAL authors note: "..if a child performs in the average range on the Verbal, Nonverbal, and Composite Indexes, but the Delayed Recall Index is more than 13 points below the other indexes, this would be significant at the $p<.05$ level. This would be an indication of disturbed retention of information in this child." (p. 42). This sort of blanket statement, if taken out of context of the manual, must be viewed with caution.

One further caution. The manual discusses how to link a subtest or composite score result with brain hemisphere pathology (pp 44-48). Although the authors caution against making specific assertions based on the test results, they then offer examples of this type of interpretation. The potential for abuse in this area seems immense.

Conclusion

The TOMAL appears to offer a new and comprehensive addition in psychologists' choice of memory assessment tools.

In the process of reviewing the test, these reviewers administered over twenty TOMAL's. It was found at first to be difficult and often burdensome but after repeated administrations increased in its user friendliness. Interpretation required more time and extreme caution since the test is new and the manual offers limited guidelines. The theoretical nature of the test is still undefined and examiners may choose from content driven, process derived, or expert derived factors.

As stated earlier the TOMAL has a steep learning curve for administration. Although this is definitely a factor in the ease of administration it is secondary to the reactions of the children we observed with this instrument. Both children and older adolescents found the test to be difficult, and for some, very challenging. Frustration was evidenced by quite a few of the subjects. One wonders how a child who is deficient in memory skills, and who is aware of such deficiencies, might react to the administration of the TOMAL. One student's comment was particularly telling. After being administered the entire TOMAL in standardized fashion, the child noted to the examiner "This test made me feel stupid." These factors need to be considered when choosing this instrument. Some examiners may wish to only use portions of this test. The information contained in the manual is insufficient to judge subtest specificity. As with any standardized test, examiners who pick and choose subtests run the risk of misinterpretation. Given that the TOMAL factor analysis did not support the hypothesized factors, examiners who simply use individual subtests, or subtest groupings must do so with the knowledge that they are reliant on clinical inference rather than factorial inference.

Concerns about the protocol, the scoring, and the interpretation are all issues that raised concern to these reviewers. Valuable research is yet to be done using the TOMAL and school aged populations so that its merit can be proven.

Ron Dumont

Peter Whelley

Rita Comtois

References

Kolb, B. & Whishaw, I.Q. (1990). Fundamentals of human neuropsychology. New York: Freeman.

Sheslow, D., & Adams, W. (1990). Wide range assessment of memory and learning. Wilmington, DE: Jastak Associates.

Squire, L. (1987). Memory and the brain. New York: Oxford University Press.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792395
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# REY-OSTERREITH COMPLEX FIGURE

**Sections:**



**Never seen the Rey-Osterreith?**

The Rey-Osterreith Complex Figure (ROCF) was devised in 1941 by the Swiss psychologist Andre Rey (cited in Lezak, 1983) for the purpose of assessing perceptual organization and visual memory in brain injured subjects. Since that time, wide use of the Rey-Osterreith Complex Figure has been reported, yet little empirical data can be found to support its use. Waber and Holmes (1986) reported that the ROCF "permits assessment of a variety of cognitive processes, including planning and organizational skills and problem-solving strategies, as well as perceptual, motor, and memory functions." (page 563). It is quick, easy, and inexpensive to administer.

Standardized procedures for the Rey-Osterreith Complex Figure were published by Osterreith in 1944 (cited in Lezak, 1983). Osterreith standardized the administration procedure, obtained normative data from 230 normal children and 60 adults, and provided interpretative guidance. Aside from Osterreith 's original publication, the ROCF is not available as a test package complete with test materials, administration procedures, detailed scoring criteria, normative data, and reliability analysis. The ROCF is passed from one evaluator to the next as two reproduced sheets of paper, one containing the design, as shown in figure 1, and the other containing an 18-element scoring criteria. The recycling of the ROCF does pose obvious risks. The most obvious is that the integrity of the design may be compromised by repeated reproduction. Waber and Holmes (1986) acknowledged that, after testing over 400 children, an extra line was discovered in the ROCF design, which they used.

## Synopsis of the Research Literature

Sources for this activity were identified through the PsycLIT Database compiled by the PsycINFO Users Services of the American Psychological Association. This database includes over 1300 journals and provides computerized access to international literature in psychology and related disciplines. The database includes publications from 1974 and is updated quarterly. The PsycLIT Database compares article titles and key words of abstracts to a search word entered by the user. For this literature review, the search word was "Osterreith". The search yielded only ten articles.

## Applications

Klicpera (1983) examined problem-solving behavior of reading disabled boys on visuomotor tasks. The ROCF was administered to 33 boys with 10s above 85 and a reading retardation of more than 1 1/2 years on standardized reading tests. A comparison group of 18 boys with no known reading disability and an average IQ of 107.2 was also tested, Klicpera (1983) found that the poor readers showed significantly poorer performance in recall on the ROCF than the control group.

Waber and Holmes (1985) formulated a new method for evaluating copy productions of the ROCF and presented normative data for 454 children from middle to lower class district ranging in age from 5 and 14. They also described developmental changes evident in the copy productions that may be of interest in neuropsychological evaluations.

Waber and Holmes (1986) provided a continuation of the above study by describing the recall productions from the same test group and prescribing a method for evaluating recall productions of the ROCF. Normative data was also provided.

Waber and Bernstein (1989) administered the ROCF to fifth and eighth graders. Within each grade, one group studied the figure visually while another copied it. Each group was tested the same on the recall task, and the results were compared. The eighth graders who studied the figure visually performed better than the group who first copied it.

Bennett-Levy (1984) examined the factors involved in performance on the ROCF among 107 adults and derived a regression equation for predicting recall performance from copy performance.

Loring, Lee, and Meador (1988) administered the ROCF to 29 patients with partial complex epileptic seizures. The seizures originated in the right temporal lobe in 15 patients and the left temporal lobe in 14 patients. No significant qualitative differences were found when applying Osterreith's scoring criteria. However, after applying a new scoring system to account for qualitative differences, the patients with right temporal lobe seizures were found to perform significantly poorer than the comparison group.

Crossen and Wiens (1988) administered the ROCF to 13 adults with moderate to severe head injuries. Recall results were reported to be below normative standards. However, the normative standards were not defined in the research report.

Levine, Warach, Benowitz, and Calvanio (1986) studied improvement and recovery factors in patients who had experienced a stroke in the right hemisphere. The ROCF was used, along with other tests, to classify severity.

Bigler (1988) administered a battery of psychometric tests including the Rey-Osterreith Complex Figure to patients with verified frontal lobe damage. The patients with left frontal lobe damage performed significantly worse than the patients with right frontal lobe damage did.

Bigler, Rosa, Shultz, Hall, and Harris (1989) administered the ROCF to 52 closed head Injury patients ranging from 18 to 55 years old and 42 Alzheimer patients ranging from 55 to 85 years old. Bigler et al. (1989) reported the performance of the Alzheimer patients on both the copy and recall tasks to be poorer than that of the head injury patients.

## Test Administration

Test administration, when described, was generally consistent with Osterreith's standardized procedure. Lezak (1983) describes the procedure as follows:

The subject is first instructed to copy the figure, which has been so set out that its length runs along the subject's horizontal plane. The examiner watches the subject's performance closely. Each time the subject completes a section of the drawing, the examiner hands him a different colored pencil and notes the order of colors.... Time to completion is recorded and both test figure and the subject's drawings are removed. This is usually followed by one or more recall trials. (P.395)

This review showed that major differences in test administration with respect to: individual versus group administration, the use of colored pencils, and delay time between copy and recall tasks. None of the researchers indicated placing time restraints on the actual conduct of the copy or recall tasks.

## Individual Versus Group Administration

The primary determinants to group or individual testing were age, and physical and mental capabilities. Researchers studying head injury (Bigler, 1988; Bigler et al., 1989; Crossen & Wiens, 1988; Levine et al., 1988; Loring et al.,1988), tested subjects individually. Researchers studying only normal children (Waber & Bernstein, 1989; Waber & Holmes, 1986; Waber & Holmes, 1987), tested children as a group with the exception of kindergarten children who were tested individually. All subjects were tested individually in Bennett-Levy's study (1984) of normal adults and Klicpera's study (1983) of older normal children and children with reading problems. In these studies, individual testing was done to enable the examiner to note procedural methods rather than because of limitations of the test group.

## Colored Pencils

Colored pencils are used to assess a subject's procedural method and are part of the original procedural method set forth by Osterreith. Changing and controlling the order in which colored pencils are used enables the examiner to determine the sequence with which each element of the drawing is completed. More importantly the examiner can determine if a part-oriented or configurational approach is used. In a configurational approach, the subject completes a framework consisting of the large rectangle and diagonals and then fills in the detail. The use of the colored pencils when administering the ROCF to a group enables the researcher to make this type of determination. For example, if a subject drew a contiguous line for a diagonal within the rectangle, this would be drawn in one color. However, it a diagonal line is completed in two segments, before and after an intersection with the other diagonal line, two colors would probably be evident.

Lezak (1983) described the use of colored pencils but also explained that some examiners keep a detailed record of the subject's copying sequence by reproducing the performance and numbering each unit in the order that it is drawn. Bigler et al. (1989) and Loring et al. (1988) reported standard administration of the ROCF was performed according to Lezak (1983). Although this is interpreted as meaning the use of colored pencils, it remains unclear as Lezak (1983) described both methods. As stated, Bennett-Levy (1984) and Klipera (1983) used one pencil and relied on examiner observation to appraise procedural method. Levine et al. (1986), Bigler (1988), and Crossen and Wiens (1988) did not specify this element of the procedure. Waber and Holmes (1985) offered a detailed description of the use of colored pencils:

When the tester signaled the children to start, they were to pick up a designated colored pencil and begin to copy the design. When told to switch, they would put down that pencil and continue drawing with the next color designated. This procedure continued until all the colors had been used.

The time limit for each color was 60s for kindergarten and first grade children, 45s for second to sixth graders, and 30s for seventh and eighth graders. The red pencil was always used last since red was thought to be a more salient color than the others, and the order of presentation of the other four colors was randomized for each classroom group. (p. 267)

## Delay Times

Delay time between copy and recall productions varied among examiners. Lezak (1983) reported awareness of examiners who used 20-minute, 30-minute, 40-minute, and 45-minute delay times. However, she also cites research that showed little difference in test results using different delay times as long as the delay is within one hour. The researchers who reported this information used the following delay times.

Klicpera (1983): Immediate and 20 minute

Bennett-Levy (1984): 40 minutes

Waber and Holmes (1986): Immediate and 20-minutes

Waber and Bernstein (1989): 20-minutes

Loring et al. (1988): 30-minutes

Bigler et al. (1989): 3-minutes

## Scoring Criteria

Scoring criteria used by the researchers reflects the greatest departure from Osterreith's test procedures. A review of Osterreith's scoring criteria is found in Lezak (1983). Osterreith defined 18 units of the drawing and assigned point values of 0 to 2 to each unit depending upon the degree to which the units are correctly drawn and placed. Osterreith evaluated organizational structure in the context of seven different procedural types. Waber and Holmes (1 985) wrote that Osterreith's criteria provides "a basis for comparing a child's performance to that of

the normal group, [but] it is insensitive to aspects of the production that may be of considerable theoretical significance. "(p. 265) Specifically, Waber and Holmes (1985) contended that Osterreith's criteria lacks "(1) a valid and reliable method for assessing parameters that are most relevant for neuropsychological diagnosis; and, (2) detailed developmental descriptive data." (p. 265)

Waber and Holmes (1985, 1986) and Bennett-Levy (1984) presented details of alternative scoring systems. To analyze detail, Waber and Holmes (1985, 1986) broke the design down into the smallest line sediments possible and objectively evaluated each segment as to accuracy, intersections, alignments, and direction of execution. Using this method, interrater reliability was calculated at 95%. Additionally, the drawings were evaluated for goodness of organization accuracy. The organization rating was based on a 5-point scale ranging from poor (1) to excellent (5). Style rating included four categories: (a) part-oriented, (b) exterior configuration/anterior part-oriented; (c) exterior part-oriented/interior configurational; and (d) configurational.

Bennett-Levy (1984) used Osterreith's 18-point scale for detail evaluation. However, a strict application of Osterreith's scale was employed on the copy production, and a lax application was employed on the recall production. Organizational structure was evaluated in terms of symmetry and good continuation rather than Osterreith's seven procedural types. Likewise, Klicpera (1983) used Osterreith's 18-point scale but developed a more detailed analysis process for evaluating organization.

Bigler (1988) and Bigler et al. (1989) scored the ROCF according to a method outlined by Denman where a maximum score of 72 may be achieved. Two references were provided in the literature: "Denman, S. B. (1984). Denman *Neuropsychology Memory Scale* [italics added], Charleston, SC: Privately published" (Bigler. 1988, p. 295), and "Denman S. B. (1984). Manual for the Denman *Neuropsychology Memory Scale* [italics added], Charleston, SC: Privately published" (Bigler et al., 1989, p. 280). No further direction as to scoring method was offered.

Levine et al. (1986) did not specified the scoring criteria used on the ROCF. These researchers combined ratings on the ROCF with ratings on other tests to develop an overall severity rating for stroke victims.

Loring et al.(1988) scored the copy production in accordance with Osterreith's 18-point detail rating only for the purpose of excluding patients from the study who did not achieve a certain score. Recall productions were also scored according to Osterreith's 18-point detail rating, but this yielded no significant differences. Loring developed a new qualitative rating focusing on distortion and misplacement errors. With this rating system, significant differences were found between right and left temporal lobe epilepsy patients, with the patients with right temporal epilepsy exhibiting a greater number of errors.

Crossen and Wiens (1 988) reported only that head injury subjects scored well below normative standards as reported by Wiens. The cited reference was identified as follows: "Wiens, A.N., McMinn, M.R. & Crossen, J.R. Rey-Osterreith Complex Figure Test: Development of norms for healthy adults. In preparation" (Crossen & Wiens, 1988, p. 399). No further direction as to scoring method was offered.

## Normative Data

In this section, the use of normative data by the researchers is described. The literature revealed several uniquely different approaches to the use of control or comparison groups. Comparison groups consisted of normal populations, another dysfunctional group, or both. The nature of the studies shed some light on the researchers' approaches, and some approaches seem to require further justification.

Osterreith obtained normative data in 1944. This normative data enabled a tester to compare a subject's performance against a comparable normal population. Quantitative copy and recall scores were compared to quantitative norms.

Procedural methods were normed as percentiles, or likelihood of occurrence. As reported by Lezak (1983), Osterreith defined the adult's average score on the copy production to be 32 and on the recall production to be 22. Yet Loring et al. (1988), who used Osterreith's 18-point detail scoring criteria, excluded subjects who scored a 34 on the copy trial "so that constructional deficits would not confound examination of memory performance" (p. 241).

Without exception, researchers reviewed herein to evaluate organizational processes did not use Osterreith's normative data. Use of Osterreith's normative data for detail analysis was implied in terms of assessing good or poor performances, but no researcher provided specific comparisons within this context.

Waber and Holmes (1985, 1986) and Bennett-Levy (1984) studied large normal populations and evaluated subjects in relation to each other. Waber and Holmes (1985, 1986) offered normative data. Loring et al. (1988) compared right and left temporal lobe epilepsy patients to each other. Crossen and Wiens (1988) used unpublished normative data and reported their test group of head injury patients to all be below normal. Levine et al. (1986) studied stroke victims and compared subjects to each other. Klicpera (1983) compared normal children to children with reading problems. Bigler et al. (1989) compared head injured patients aged 18 to 55 to Alzheimer's patients aged 55 to 85. In another study, Bigler (1988) compared two left and two right frontal lobe damaged patients to each other.

## Research Results

Waber and Holmes' 1985 study of the ROCF copy task was intended to provide a reliable and valid method of evaluating the ROCF, and to describe cognitive developmental difference among children of different ages. The derived method and resultant normative data is provided in the research report. High interrater statistics were reported for all facets of the evaluation process. Significant developmental features were reported as follows:

1. Nearly total accuracy was achieved by age nine, with little change occurring thereafter.

2. Young children, between the ages of 5 and 7, were equally likely to start on the left or right side. At age 8, a left-side preference was shown by 64% of the children. The proportion increased to 80% between ages 9 and 12, and by age 13, reached 90%. Directional preferences were viewed as diagnostically significant only at age 9 or older. Direction of execution was not related to handedness.

3. Organization was evaluated as configurational or part-oriented. Copies became more configurational with age. Children who copied from right to left produced more part- oriented productions than those who drew from left to right. Part-oriented drawings were seen more often in left-handed children. No significant difference was seen between sexes.

4. The base rectangle was the salient organizational unit used in configurational drawings. When children began with the base rectangle, information was treated more logically than figurally. This observation was deemed consistent with Piaget's cognitive developmental theory describing the evolution from concrete to logical reasoning. By age 13, logical thinking

dominates.

Waber and Holmes' 1986 research report describes the recall performance of the same study group used in their 1985 study. The goals of this study were to devise a reliable and valid evaluation criteria and describe developmental processes. The evaluation criteria and normative data is provided in the research report.

Developmental features demonstrated by the recall task were as follows:

1. Accuracy increased with the more configurational versus part-oriented approach. No child achieved total accuracy.

2. The base rectangle and main structure was recalled almost perfectly from age 9 onwards.

3. Material on the left side of the design was recalled better than that on the right up to age 8, at which point recall for the two sides was equivalent.

4. No significance was realized in the immediate versus delayed (20 minutes) recall for configurational items. Delayed recall resulted in further loss of internal detail.

5. Recall productions were more configurational than copy productions except among the youngest children.

6. The delayed recall resulted in a marked shift to a more configurational approach as compared to the immediate recall.

7. Organizational style displayed on the copy production was strongly correlated to the recall production.

8. Among five-year-olds, part-oriented and configural approaches were observed equally.

9. Part-orientation in memory productions was rarely demonstrated beyond age 9.

Waber and Bernstein's 1989 study provided greater insight into configurational versus part-oriented approaches in cognitive development. Fifth and eighth graders were tested in two groups each. One group performed the copy task of the ROCF while the other group studied the design visually. Then all children performed the recall task in the, same manner. The conclusions presented by the researchers are summarized as follows:

1. Preadolescent children who visually studied the figure produced more configurational recall productions than did those who first copied the figure. This effect was more pronounced among boys.

2. Preadolescent children who visually studied the figure produced better-organized recall productions and were more accurate in reproducing the structural components when compared to the copy group. There was, however, no group difference for retention of detail.

3. Eighth graders showed a prevalence of configurational approach and increased organizational skill independent of modality of input. In summarizing this result, the researchers found that the motor program comes to be dominated by the organizational information carried by the visual code.

4. Additionally, Waber and Bernstein (1989) reported: "... the most provocative finding is that among preadolescent children, motor input apparently *interfered* [italics added] with, rather than supported, efficient encoding of visuospatial information. Eliminating motor input in the encoding phase enhanced visual memory, with fifth graders performing like eighth graders on all parameters measured under these conditions. The fact that the motor output entailed in the immediate recall did not lead to a more part-oriented approach in either that condition or the delayed recall among the visual group further localizes the phenomenon to the encoding phase. Once the material is visually encoded, the motor code can no longer preempt it. (P. 13)

Bennett-Levy (1984) developed a means of predicting recall performance. The derivation of a regression equation is included in the research report. The following observations regarding a normal adult population were offered:

1. Copy strategy, rated in terms of Gestalt concepts of symmetry and good continuation, was the primary determinant of recall score.

2. Estimated 10 (reading ability) was significantly correlated with copy and recall performance, but strategy effects were shown to be wholly independent of IQ.

3. Copy time was a determinant to copy score but not recall score.

4. Age was correlated to both copy and recall performance among early and middle adulthood subjects. This was reported as a surprising result since studies reviewed by the researchers had suggested that age in the early to middle adult range should not make a difference.

Loring et al. (1988) reported no significant differences in mean scores on the ROCF between subjects with right temporal lobe epilepsy and left temporal lobe epilepsy when using standard scoring criteria (undefined). However, Loring et al. (1988) developed a qualitative rating that did yield significant differences. Then, right temporal lobe epilepsy subjects exhibited a significantly greater number of errors. Loring et al. (1988) wrote:

"The present report illustrated standard CF [Complex Figure] scoring criteria are inadequate to characterize the types of errors observed in patients with right TLE [temporal lobe epilepsy]. .... the type of responses that we observe with right hemisphere seizure focus involve distortion or misplacement. Obviously, a scoring system that scores principally for presence or absence of elements, with little or secondary weight to misplacement, cannot adequately capture the quality of errors..." (p. 244 & 245).

Klicpera (1983) reported lower recall scores for the children classified as poor readers as compared to children with no known learning difficulty. This was attributed to poorer planning by the poor readers during the copy phase. The poor readers were less likely than the children in the control group to draw the structure of the design first and then fill in the detail. Rather, they began to reproduce details much

sooner than controls. The organizational scheme used when copying was also the scheme used during recall. Recall scores were worse for poor readers even when perfect scores were realized for copy productions. This was attributed to the organizational scheme employed. Klicpera (1983) concluded that the "study suggest[s] that dyslexic children have a developmental delay in an area that is broader than purely verbal skills" (p. 80).

It is difficult or impossible to isolate significant ROCF test result data from the research of Crossen and Wiens (1988), Levine et al. (1986), Bigler et al. (1989), and Bigler (1988). Crossen and Wiens (1988) reported ROCF test results "well below normative standards" (p. 395) for subjects with head injury but provided no discrepancy analysis for the ROCF in particular. Levine et al. (1986) offered test results from the study of stroke victims which were based on a composite of psychometric and medical data. Bigler et al. (1989) compared Alzheimer subjects with closed head injury subjects resulting in large age differences between groups. Alzheimer patients performed more poorly on the ROCF, but the analysis explaining this is not very convincing since the age factor is not given adequate attention.

Bigler's 1988 test group consisted of only four frontal lobe damaged patients, two with right frontal lobe damage and two with left frontal lobe damage. When reviewed against the other literature, one would expect poorer performance on the ROCF from the right frontal lobe damaged test groups as the right hemisphere enables configurational organizational abilities and the left hemisphere is prominent in part-oriented organizational approaches. However, Bigler (1988) did not support this. In fact, the two left frontal lobe damaged subjects, for whom the right hemisphere was undamaged, demonstrated much poorer performances on the recall portion of the ROCF. The copy portion was poor for just one of the left frontal lobe damaged subjects. Bigler (1988) explained the left frontal lobe damaged subject who had a good copy score and a poor recall score this way: " ...because he did not anticipate that he was going to be required to recall the figure from memory he did not develop any effective strategies for recall while he was copying." (p. 294). This explanation is inadequate, as none of the subjects could have anticipated the recall task. The small number of subjects evaluated makes any conclusions from this study weak.

## Discussion

The purpose of this literature review was to define current application of the Rey-Osterreith Complex Figure. Although wide use of the ROCF was reported or implied in the literature, the limited number of published research articles on the subject makes this questionable and prompts the need for a survey across a broad spectrum of test users (i.e. neuropsychologists, clinical psychologists, school psychologists).

Test administration, scoring criteria, the use of normative data, and test results were specifically addressed in this review. Examination of these elements showed significant deviation not only from the standardized procedures set forth by Osterreith, but from one researcher to another. This deviation is appropriate with regards to normative data, but it is an area of concern with regards to other aspects of the test. For example, evaluation of test results must include consideration of the copy score derived from the elements of the ROCF design within the context of the organizational approach. Consideration must also be given to the subject's age as the research indicated a correlation between age and score and also implied that it may not be appropriate to administer the ROCF to children under 9 years old.

The literature implies a very broad application of the ROCF. Children and adults were tested as well as normal, learning disabled, and brain damaged subjects. The literature also showed application in neuropsychology and education. In neuropsychology, the ROCF may be used to localize and assess the magnitude of brain damage. In education, the ROCF may be used to evaluate input processing of a child suspected of having a learning disability.

There is a clear need for users of the ROCF to move towards a standardized method of administering and evaluating performance on the ROCF. Such a movement would also require the development of current normative data from normal, brain-damaged, and learning disabled children and adults. The natural conclusion of such an effort would be to revise Osterreith's standardized procedures and publish the revision in the form of a user's manual for the ROCF.

## REFERENCES


Bennett-Levy, J. (1984). Determinants of performance on the Rey-Osterreith Complex Figure Test: An analysis and a new technique for single case assessment. British Journal of Clinical Psychology, 23, 109-119

Bigler, E. D. (1956). Frontal lobe damage and neuropsychological assessment. Archives of Clinical Neuropsychology 3, 279 - 297.

Bigler, E. D., Rose, L., Schultz, F., Hall, S., and Harris, J. (1989). Rey-Auditory Verbal Learning and Rey-Osterreith Complex Figure design Performance in Alzheimer's Disease and closed head injury Journal of Clinical Psychology, 45 (2), 277-280.

Crosson, John R. & Wiens, Arthur N. (1 956). Residual Neuropsychological deficits following head-injury on the Wechsler Memory ScaIe - Revised. The Clinical Neuropsychologist, 2 (4), 393-399.

Klicpera, C. (I 983). Poor planning as a characteristic of problem-solving behavior in dyslexic children. Acta Paedopsychiatrica, 49 (1/2), 73-82.

Levine, D. N., Warach, J. D., Benowitz, L., and Calvanio, R.(1986). Left spatial neglect: Effects of lesion size and premorbid brain atrophy on severity and recovery following right cerebral interaction. Neurology, 36,362-366.

Lezak, Muriel D. (1963). Neuropsychological Assessment 2nd Edition. New York: Oxford University Press.

Loring, D. W., Lee, G. P., & Meador, K. J. (1988). Revising the Rey-Osterreith: Rating right hemisphere recall. Archives of Clinical Neuropsychology, 3, 239-247.

Waber, D. P. & Bernstein, J H. (1989). Remembering the Rey-Osterreith Complex Figure. A dual-code cognitive neuropsychological model. Developmental Neuropsychology, 5 (1), 1- 15.

Waber, D. P. & Holmes, J M. (1986). Assessing children's memory productions of Rey-Osterreith Complex Figure. Journal of Clinical and Experimental Neuropsychology, (5), 563-580

Waber, D. P. & Holmes, J M. (1986). (1985). Assessing children's productions of the Rey-Osterreith Complex Figure. Journal of Clinical and Experimental Neuropsychology, (3), 264-280.

## Normative Addendum

Bernstein, J. H. & Waber, D. P. (1996). Developmental scoring system for the Rey-Osterreith Complex Figure: Professional manual. Odessa, FL: Psychological Assessment Resources, Inc.

> Bernstein & Waber (1996) used a colored-pencil administration system with copy, immediate recall, and 15- to 20-minute delayed recall "without warning" and with "several interpolated verbal tasks" (p. 5). They have complex, meticulously detailed scoring rules for an Organization Score, Style Rating, Accuracy Scores, and Error Scores. Norms, sadly, are based on "454 children. . . . from a middle- to lower-middle-class school district in the northeastern United States. . . . from 5 through 14 years."

Kolb, B., & Whishaw, I. Q. (1985). Fundamentals of Human Neuropsychology. New York: W. H. Freeman & Co.

> Kolb & Whishaw (1985, pp. 734-736) offer norms, apparently using Lezak's (1983) choices of administration procedures and scoring criteria, based on a "randomly selected sample of [2,740] school-aged [6 through 18] subjects collected from the Lethbridge [Alberta] public and separate school systems." The considerable virtue of the large sample is offset by the single location and the lack of precision in rules for administration and scoring. Kolb & Whishaw (1985, p. 738) also offer norms, based on 2,665 people (2,357 children), for the Draw-a-bicycle test.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792397 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# TEST REVIEW

# Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA) (1994)

Carole L. Cruse, Ron Dumont, & John Willis

**AUTHORS**: Richard Woodcock, Kevin McGrew, & Judy Werder.

**PUBLISHER**: Riverside Publishing. $138 for the test easel with manual, 25 record forms with subject worksheets, and 5.25" and 3.5" computer disks (for either IBM- compatible or Apple/Macintosh). The Mini-Battery's scoring is done solely by computer, using the MBA Scoring and Reporting Computer Program. The manual does not contain any tables or charts of normative data.

**PURPOSE**: The Woodcock-McGrew-Werder Mini-Battery of Achievement (MBA) is an approximately 30-minute, one-easel assessment of reading, math, writing, and knowledge intended for various educational "screening" purposes for ages 4 to 95.

**STANDARDIZATION**: The Manual refers the reader to the WJ-R Technical Manual for information regarding the MBA norming since the norms of the MBA are based on a common sample with the WJ-R. The MBA is an amalgamation of many existing WJ-R items and alternates that were tested, but not included in the WJ-R. The MBA consists of 269 items, with 101 items (38%) taken from the WJ-R, Form A, and 119 items (44%) taken from Form B.

**RELIABILITY**: Split-half reliabilities based on the entire norming sample at ages 5, 6, 9, 13, 18, and four adult groups ranged from .92 to .94 for the three Basic Skills subtests. Factual Knowledge has a median reliability of .87, while the Cluster reliability is .93. Test-retest reliabilities range from .85 to .94 for subtests and from .94 to .97 for the Basic Skills Cluster.

**VALIDITY**: Content validity is discussed as it relates to the content of the items used. Items were selected based on "item validity studies as well as expert opinion". Concurrent validity was established using the same samples used for the test-retest studies. Even with the substantial likeness in origin, form, and content, this did not result in remarkably higher correlations between the MBA and the WJ-R and other achievement tests (PIAT, K-BIT, WRAT). Patterns of subtest intercorrelations provide data for construct validity. The median correlations ranged from .65 to .80, indicating significant relationships among the subtests.

**ADMINISTRATION AND SCORING**: The test consists of four main sections: Reading, Writing, Mathematics, and Factual Knowledge. The Reading section asks the examinee to identify letters and words, state antonyms for words, and answer questions based on paragraphs read. The Writing section asks the examinee to print letters, words, punctuation marks etc. when the task is read aloud by the examiner. The examinee is also asked to identify an error in written sentences or passages, which may be due to spelling, punctuation, or usage. The

Mathematics section contains subtests assessing both computation and math reasoning skills. Factual knowledge, though not included in the in the Basic Skills Cluster, assesses knowledge of science, social studies, art, music, and literature.

Starting points are "suggested" for different school-based levels. Response time has no defined parameters, although examiners are encouraged not to spend "unnecessary time" before advancing on to the next item. The MBA is advertised as a test that can be administered by paraprofessionals because of its ease of administration and scoring. Basal and ceiling rules are given, but they may seem confusing to paraprofessionals. The basal rule states "If the 4 lowest numbered items given are not all correct, return to the starting point. Then test backward, full page by full page, until the 4 lowest numbered items given are all correct...." The inexperienced examiner may assume that they have completed the testing for the subtest. There is no reminder to return to the highest item and continue to test until the discontinue criterion is met. Scoring guidelines are generally very clear. The MBA Scoring and Reporting Computer Program supplies derived scores for the total Reading, Writing, Mathematics, Basic Skills, and Factual Knowledge measures. No scores are reported for the separate subtests that make up the measures. A potential interpretive problem may result if a child has a strength in one skill that is significant enough to overshadow a weakness in another skill within the same academic domain. The resulting total "averaged" score may be simply a misleading artifact that has hidden the strength and/or weakness.

The Scoring and Reporting Program includes, and emphasizes, age- or grade-equivalent scores in both the statistical table and the narrative section of the report printout. The statistical table includes: age- or grade-equivalent scores, percentile rank, standard score, SEm for standard scores, Normal Curve Equivalent, and T-score. Scores are reported as single scores, with no confidence intervals.

**PRACTICAL CONSIDERATIONS**: These reviewers administered the MBA to a range of students and adults as part of this review. Testing time varied with age, with only the very young children, 4 to 6 years, meeting the 25-minute to half-hour time criterion stated in the manual. Younger children were unable to perform the tasks asked of them on at least four of the subtests (i.e., Reading Part B: Vocabulary, Part C: Comprehension, Writing Part B: Dictation, and Mathematics Part A: Calculation), causing quick discontinuation and greatly lessening the testing time. Unfortunately, that limitation also effects the utility of the resulting scores. The time estimation for the test was not met by the older children and adults. Skill level change between items was sudden and drastic, making some of the examinees we tested feel frustrated and ineffectual.

These reviewers were concerned about the obtainable scores on the MBA. A four-year-old child, obtaining a raw score of 0 on every subtest, would receive standard scores ranging from <18 to <108. This same child, obtaining a raw score of 1 on each subtest, would receive standard scores ranging from 25 to 152! The MBA, although normed for children as young as four, supplies little useful or meaningful information at the low age levels.

**CONCLUSIONS**: It is difficult to consider the MBA a unique test in its own right, when approximately 82% of its items are taken from the WJ-R Achievement Tests. It is more like a short form derived from a complete test. If the WJ-R were used for follow-up testing when indicated by MBA results, it is possible that the student will be given up to approximately 101 or 119 items already taken in the MBA.

The computer program was straight-forward and easy to use. Data entry takes typically less than 2 minutes. The program allows results to be either printed or shown on the screen. While the computer scoring capabilities of the MBA are both timely and impressive, including a scoring disk without a normative manual is not practical. Having the capacity to double check computer obtained scores is imperative when making assessment decisions.

Overall, the MBA is a step above other brief screening tests because it assesses a broader range of academic skills. The MBA is a good achievement screener since it is easy to give and understand and has the potential to elicit meaningful results in a timely manner. Examiners may be tempted to use the MBA as a "quick and dirty" substitute for the WJ-R, because of its brief testing time, extensive scope, and a quick report. Examiners must use the test solely for the purposes expressly stated by the authors.

**REFERENCES**:

Woodcock, R. W., & Johnson, M. B., (1989). Woodcock-Johnson Psychoeducational Battery-Revised. Chicago: Riverside Publishing.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792398 HITOMETER

Dunmont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Short Form Assessments: Some thoughts by Cisco and Eggbert

John and I have been asked a number of time over the past year about our thoughts on the use of "Short Forms" for the WISC-III. Do they exist? How valid and reliable are they? Are there any easy charts or tables for creating them? Should they be used? Would we be willing to provide charts and tables for them? Because I am somewhat biased about short forms, having recently published an article about a particular WISC-III combination in Psychology in the Schools, John agreed to disagree with me on this article. (As many of you know, John and I have argued with each other, as well as occasionally agreed, at workshops a number of times and have affectionately been labeled, and have subsequently adopted as our moniker, the nicknames "Cisco and Eggbert" of testing.

"Eggbert, have you got any thoughts on short forms IQ tests? We, or should I say "I", always seem to be complaining about the amount of time we spend testing and evaluating children. Wouldn't some short form be a useful time saver? The usual purpose for readministering intelligence tests to children already identified as learning disabled is usually to simply validate the individual's cognitive functioning as being at least "normal". This having been done, why do it again? Studies have found the results of IQ tests do not usually contribute to recommendations regarding remediation or intervention. They're often simply used to make decisions regarding the continuation of the handicapping condition. Wouldn't these limitation, coupled with the amount of time that we spend administering lengthy assessments, make us want to explore alternative procedures for validating a student's cognitive functioning during a three year evaluation? Don't we test too much and without reason?"

"Cisco, you know me, of course I have some thoughts about short form IQ tests. Since you researched short forms and then created and validated one, let me say that yours was the right way to do it, and doing it any less thoroughly would be reckless. However, there is a right way to fillet a guppy for dinner or to perform brain surgery with only two instruments, and if you are going to do either one you should do it right...but why would you be doing them? If the studies you referred to are right, and we aren't testing to discover current patterns of intellectual abilities or in possible changes in patterns, what are we interested in? Do we simply want a number?"

"Easy Eggbert, you're getting carried away. Special education teams deal with numbers every day. It's the nature of the beast. Some teams 'require' the re-assessment of a child's intellect every three years. That itself seems silly and may be the problem, but since they do, it turns into a question of the redundancy of the testing. We all know places where the motto is 'If it moves..WISC it.' How many times have we tested kids who know the test better than we do because they've been given a WISC 4 times. They ought to be giving the test to us and at times they know the material enough that they are answering before we ask. I love giving the DAS on reevaluation and really stumping the kids who are expecting a WISC. If the child is already identified as learning disabled, wouldn't assessing IQ using a short form be beneficial. Less time is taken on something that we probably won't use anyway, and more emphasis can be placed on areas of identified weakness. And maybe, philosophically, shouldn't we be de-emphasizing the whole idea of IQ as an exclusionary factor in our assessments?"

"Great idea but will it happen? Think about the meetings we've been in where the IQ has changed over a three year interval. There are few sights more pathetic than a team scrambling around trying to explain away a drop in IQ score on re-evaluation. Teams hesitate to admit that three years in a wonderfully created individualized educational setting has cost a child a significant number of IQ points. Teams also don't seem to know what to do with increases in IQ scores. Does this mean that the child has actually gotten smarter, but is learning even less in the special education program? Most of the time, the teams simply explain away unexpected IQ scores and pay no attention to expected ones. If we ignore the retest, why bother in the first place? And what of the dilemma for the teams that do take all the numbers seriously? If a learning disabled child's IQ score drops toward the already-low achievement scores, will the team decide the child is no longer disabled? The child used to have a high IQ with low achievement and was considered educationally challenged by a learning disability. Now the achievement is still low but the IQ has dropped as well, so the child is no longer disabled and no longer in need of service. Decisions like these make Joseph Heller's Catch-22 look like a triumph of rationality. "

"We agree on all this, Eggbert. I think! The testing as it stands is problematic because of the way we use them. But if the team requests an intellectual assessment, and the chances are that nothing more than the total score will be used, and you already have a complete valid assessment, I suggest that the use of a "good" short form is justified. Clearly some rationale for the choice of the short form is necessary, as well as determination of reliability, validity, and standard error is needed. I am not advocating IQ roulette. Examiners choosing or creating short forms must understand the properties of those forms."

"Yes, and the operative word is 'good'. If you do use a short form remember: a short form of a well constructed test is better than a short test with known shortcomings; three or more subtests tapping different abilities is better than a short test tapping only one or two; and it's clearly a professional, ethical, and legal requirement to use an instrument with specified reliability. Your short form procedure meets these requirements, whereas brief tests do not, and estimating, averaging, and prorating certainly do not."

Having said all that, for anyone interested in a short paper on how to use the Tellegen & Briggs formulas to create and validate short forms of major tests, use this link.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792399 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Steps in Creating Short Forms of IQ Tests

## (Quick, Dirty, and Not Recommended)

In order to develop the statistics for a short form, the reliability coefficient and intercorrelation of each subtest must be determined.

**Step 1: Choose a Short Form**

Create a table that lists the reliability of each chosen subtest along with the correlation matrix for those same subtests. For example, suppose you want a five-subtest short form for the WISC-III that uses Information (I), Vocabulary (V), Picture Completion (PC), Coding (Cd), and Block Design (BD) [the short form developed by Dumont and Faro]. The table below was created adapting information provided in the WISC-III manual. (The relevant columns are numbered 1 through 6 and the relevant rows are lettered A through F.)

| | 1 | | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| | r | | | | | | |
| A | .84 | I | I | V | PC | Cd | |
| B | .87 | V | .70 | | | | |
| C | .77 | PC | .47 | .45 | | | |
| D | .79 | Cd | .21 | .26 | .18 | | |
| E | .87 | BD | .48 | .46 | .52 | .27 | |
| F | 4.14 | | 1.86 | 1.17 | .70 | .27 | 4.00 |

**Step 2: Determine the Sum of Reliabilities**

Add the reliabilities in column 1 (rows A through E) to obtain the sum of the reliabilities. In this case the sum is equal to 4.14 (Column 1, row F).

**Step 3: Determine the Sum of Intercorrelations**

Add all of the intercorrelations. In this case they are equal to 4.00 (Column 6, row F).

**Step 4: Determine the Multiplier and Additive**

Using the number obtained in Step 3, obtain the Multiplier and the Additive for your subtest combinations from the appropriate tables linked below.

Table for Two-subtest short form

Table for Three-subtest short form

Table for Four-subtest short form

Table for Five-subtest short form

In the example above, we are making a five-subtest short form, so I would go to the Table for Five-subtest short form and look up the numbers associated with the value of 4.00. I find the correct Multiplier and Additive in the section that looks like:

| | | |
|---|---|---|
| 3.79-3.96 | 1.4 | 30 |
| 3.97-4.15 | 1.4 | 31 |
| 4.16-4.35 | 1.4 | 32 |
| 4.36-4.56 | 1.3 | 35 |

The numbers are 1.4 and 31. To determine a deviation quotient for the proposed short form, add together the subtest scaled scores, multiply that sum by 1.4, and finally add 31 to the product.

In our example, substituting subtest scaled scores, the five subtests Short form would be:

(I+V+PC+Cd+BD)*1.4+31
(10+9+10+10+5)*1.4+31
(44)*1.4+31
61.6+31
Deviation Quotient = 92.6

**Step 5: Determine the reliability of the short form by using the following formula**

Reliability=

$$\frac{\text{Sum of reliabilities} + 2\,(\text{Sum of Intercorrelations})}{\text{Number of Subtests} + 2\,(\text{Sum of Intercorrelations})}$$

For our example the reliability would be:

$$\frac{4.41 + 2(4.0)}{5 + 2(4.0)}$$

The reliability is .93.

**Step 6: Determine the Standard Error of Measurement**

Standard error of Measurement = Confidence level * Standard Deviation * Square Root (1-Reliability)

Confidence levels:

1.64 = 90%
1.96 = 95%

For our example:

1.96*15*Sqrt(1-.93)
SEm=7.6

---

• **Dumont, R. & Faro, C. (1993) WISC-III Short Form for Learning Disabled Children.** ***Psychology in the Schools., 30, 212-219***

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792400 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# As the Block Turns: Interpretation of block rotations on block design subtests

Ron Dumont & Andrea Mariani

**This article was first published in the NASP Communiqué**

A letter sent to the editor raised a very interesting and important question that school psychologists are frequently confronted with:

> "I have been trying to find out how to interpret a very specific incident that occurs frequently on the WISC as well as on other IQ tests. .... My question is, what does it mean when a child rotates the blocks during the Block Design subtest? Is it solely an indication of visual perception problems? Also why are there different criteria for scoring the degree of the rotations"

Block Design type tasks have been common on intelligence tests since at least 1923 when Kohs incorporated them. The Wechsler Scales' Block Design, the Stanford Binet 4th Edition's (SB-4th) Pattern Analysis, The Differential Ability Scales' (DAS) Pattern Construction, the Kaufman Assessment Battery for Children's (K-ABC) Triangles, and the Kaufman Adolescent and Adult Intelligence Test's (KAIT) Memory for Block Design are all varying forms of the block design task. Each has slightly different scoring and timing procedures. For example, the Wechsler Scales incorporate red and white plastic blocks, have strict time limits, and penalize rotations greater than 30°; the DAS, using black and yellow blocks, allow the examiner to score with and without time limits; the SB-4th, with its black and white blocks, penalizes rotations of 90°; the K-ABC, using yellow and blue foam boards, does not penalize rotations at all; and the KAIT, with large black and yellow wooden blocks, has lenient time limits and a tray provided for the examinee to construct the design in. With all the variations of the task, how is one to interpret the rotation and reversals sometimes encountered?

These authors contacted, by letter or phone, the following test experts: Dr. Aurelio Prifitera, Director, Psychological Measurement, Vice President, The Psychological Corporation; Dr. Jerome Sattler, Professor of psychology at San Diego State University and co-author of the SB-4th; Dr Colin Elliott, author of the DAS and project manager for the KAIT; and Dr Alan Kaufman, Research Professor, University of Alabama and the co-author of the KABC and the KAIT. (In this article, all quotes, unless otherwise cited, are from personal communications with these authors.)

Each of these experts were consistent in at least one area: that extreme care and caution needs to be applied when making any specific interpretation. Each generalized assumption must be tempered with an understanding of the individual being tested. As Dr. Sattler noted, examiners must keep in mind that "...no one error should be interpreted without reference to the entire performance." This is clearly a tenet that most school psychologists hopefully understand and follow. Beyond that caution, there are some possibilities as to what a rotation or a reversal might imply.

**Pathognomonic possibilities for rotations and reversals**

From an historical perspective, Dr. Prifitera noted that "Wechsler originally hypothesized that rotations were associated with reading disabilities in children.

Drs. Prifitera and Sattler briefly addressed the association between rotation errors and pathology. Dr. Sattler wrote, "It could mean a possible brain injury...". Dr. Prifitera, in his response, went even further in saying that alignment errors, "...are often associated with right hemisphere dysfunction...". Consistent with this statement is Kaplan, et. al.'s (1991) research finding that patients with right frontal brain damage commonly rotate block designs. However, also cited by Kaplan are studies which implicate left hemisphere dysfunction as the culprit in single-block rotation errors (Delis, et. al., 1986, 1988). This example of the existing dichotomy in the research presents reason to decide with care what it is that Block Design rotations signify. The overall consensus on the neurological implications of block rotations seems inconclusive. Various studies, cited in Kaplan, et. al.'s (1991) suggests that a low score on the subtest could mean that the left, right or both hemispheres are damaged.

Since the research data cited above was specific to adults and the WAIS-R, these authors ran a literature search using the PsychLit data base to retrieve articles specific to children's block rotations. Searching all articles from January 1987 to June 1994, and using combinations of terms (WISC, Block Design, rotation, error, analysis, perception, performance, implication, neurological, etc.), no direct references to studies involving children were found. This same search, substituting the term WAIS for WISC resulted in a larger number of articles. This data base search suggests a paucity of data relevant to interpretation of rotations for children. Can we assume that the interpretations relevant for an adult would be equally relevant for a child?. The reminder of Reschly and Gresham (1989) seems relevant on this point

"Neuropsychological explanations of common learning problems are based on studies of highly selected and often, extraordinarily rare individuals, and then generalized to students whose developmental and neurological status are clearly different from persons included in the basic research. These generalizations often involve inferences from persons with definite brain damage to children who have no identifiable brain injury."

**Non-pathognomonic reasons for rotations**

Although rotations can be related to brain pathology, other explanations are always possible. The reason for a child rotating a block design may be as simple as "a wild guess", as Dr. Sattler put it. Dr. Kaufman, in his letter noted a very interesting observation from the standardization of the K-ABC. When he and his colleagues tested children during pilot studies on the Triangles subtest, they often questioned children who rotated designs as to why. The children answered, "So you could see the design better", or "So it would be easier for you to know if I got it right". Dr. Elliott concurred with this finding, noting "For young children, they just may not tune into the requirements for keeping things in proper orientation. Do they even perceive that it matters? Particularly on tests that do not have the examiner seated in the same orientations as the child, the child may rotate to please the examiner."

The developmental model also holds a good argument for rotations possibly being normal. Kaufman, in his explanation of why the K-ABC does not penalize for rotations, wrote that children 4 to 5 years of age may not be expected to attend to these features. At ages 6 to 8, rotations are more meaningful, however, there is still research that indicates that, "...it is normal for some children to reverse letters and to be unable to answer reversal or rotational items correctly." Sattler, in his letter to these authors, also mentioned this possibility of perceptual immaturity. Therefore, one should take care in considering the child's overall development and consider the possibility that a rotation may not have any

negative meaning.

**Why tests penalize differently**

Regarding the rationale for judging varying degrees of rotations as an error or not, there was no one reason cited by the experts. Dr. Sattler noted that "There is no answer to why some tests penalize for 30 degrees and 90 degrees, or not at all. This is a judgment made by test constructors." Dr. Prifitera notes that "Wechsler used the 30 degree criteria for consistency and standardization in scoring." Dr. Kaufman summarized the rationale for not scoring rotations on the K-ABC by noting "We did not feel confident that rotations should always be interpreted as potential for school-age children, and we were quite sure that rotations by children ages 4 to 6 or 7 (or even 8) may have no negative meaning at all. Consequently, we believed that the best way to handle the situation was to penalize no one for rotations."

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792401 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# School Psychologists or Soothsayer?

Ron Dumont & Rob Finn

**This article was first published in the NASP Communiqué**

This past month, the editors received a letter from a school psychologist asking some questions that seems fairly common in evaluations. The questions are often raised by team members and parents and the school psychologist is delegated to answer them.

> "When I review the test scores at team meetings, sometimes I report specific weaknesses significantly below average on each scale, Verbal and Performance. Teachers and parents alike will then ask how do these specific weaknesses impact academic skill areas such as Reading, Writing, and Arithmetic. Do we know how specific low subtest scores on either Verbal or Performance scales singularly or in combination imply that there will be adverse effects on academic performance.....Should I stick to saying the FSIQ is the best predictor of academic performance and ignore 1 or 2 low subtest scores or is there data to support how low subtest scores can mean trouble for an academic area?"

This is a complex question dealing with a number of important issues. It would be impossible to try to answer the questions with a simple yes or no response. We will attempt an answer to the question by addressing the differing levels of interpretation that are alluded to in this letter. These include subtest, subtest groupings (clusters) and composite levels of interpretation.

The question of how do specific subtest weaknesses impact academic skill areas is not as straightforward as it may sound. Depending on the subtest, a low score may be tapping into one component of a particular complex skill area, but it may also be nothing more than a statistical anomaly.

Subtest level interpretation is typically thought to be the least reliable and valid strategy (Kaufman, 1979 Sattler, 1988, Elliott, 1990). When interpreting any test, it is important that those interpretations be based upon the most reliable aspects of the test. Interpretation should focus on the most general and reliable of scores and these are typically derived from the entire test (Full Scale IQ for example). Below these measures come the Index or Cluster scores, followed by shared ability factors, and finally the individual subtests. Although subtest scores are related, they differ in item content and test administration and thus these differences cause the subtest scores to vary. Subtests can, and do, differ from each other. Before one can evaluate the differences between what appear to be high or low subtest scores, one must evaluate whether these apparent differences are enough to warrant interpretation. To do so we must know if the difference is large, reliable, and significant. In statistical terms, each subtest carries with it components of shared common variance, while at the same time most also have some proportion of specific, reliable variance. Before attempting individual subtest interpretation, one must be sure that the subtest being interpreted has adequate specific variance. For example, on the WISC-III, Object Assembly has much common variance but little specific variance and probably should not be interpreted in isolation. Subtest level strength and weaknesses must be interpreted cautiously not only because of this

low specificity but also because variations (strengths or weaknesses) are common. Kaufman noted that for the Wechsler scales it was very common for children to have rather large intersubtest variability which produced the peaks and valleys in scaled scores that often serve as interpretive points.

A second caution about predicting academic performance on individual subtest strengths and weaknesses is that academic skill areas involve multiple cognitive processes working in parallel. Reading, for example, is a complex process involving basic skills, conceptual understanding, and cognitive strategies. This can be further broken down into the knowledge about letters, phonemes, morphemes, words, ideas, schema, and subject matter as well as decoding, literal comprehension, inferential comprehension, and comprehension monitoring. It becomes clear that a particular low subtests score may only scratch the surface when it comes to fully understanding its academic implications. It must also be remembered that variations between and within the different functions do occur as a result of the individual's uniqueness and therefore these variations may simply be describing that uniqueness and not necessarily any 'difficulty.' This is not to say that strengths and weaknesses are without value, but rather that identification of a learning problem must be made on an individual basis. Subtest analysis is obviously only one piece in the assessment pie, and probably not the best piece at that.

Does the use of composites or 'profiles' increase ones ability to predict learning difficulties and thus academic performance? The use of composites is certainly more statistically reliable than individual subtests, but the cautions about psychometric properties remains applicable. Before a prediction can be made about a person based on the relative value of a composite, one needs to know the reliability and integrity of such a composite. As an example, the WISC-III Freedom from Distractibility index has a high reliability coefficient yet accounts for only 3-4% of the test's common variance. Secondly, before interpreting this factor, one must be sure that it is meaningful; that the subtest scores that make it up have measured a similar skill. If the Arithmetic and Digit Span subtest scores differ by 5 points, the interpretive utility of this factor is lost.

Do certain profiles exist that can be used to identify/predict learning difficulties? The continuing scholarly debate about this issue seems only to add confusion to what we do. The answer to this question seems to depend on who you believe and the approach used to justify the analysis. For example, Kavale and Forness argued against the utility of profile analysis in their article "Meta-analysis of WISC-R Profiles-Patterns or Parodies." (1984). Their analysis of Wechsler scale data from 9,372 learning disabled children failed to distinguish these children from their normal peers on any of the ability patterns that have conventionally been held to characterize LD children's test performance. This was followed by Lawson and Inglis' "Micro-interpretation or Misinterpretation? A reply to Forness, Kavale, and Nihira." (1987). Lawson and Inglis argued that their learning disability index did reveal a pattern that distinguishes LD from a normal sample. More recently, Keith, et. al.(1992), added "Profile Analysis with the Wechsler Scales: Patterns, not Parodies". In this paper, the authors disagree with the conclusions of Kavale and Forness and offer the opinion that "proper analysis of the data reported in their article reveals that many such profiles and recategorizations are indeed significantly different for the two groups."

Is the FSIQ the best predictor of academic performance? To answer this, one might ask "Best in comparison to what? Better than the Verbal and/or the Performance IQs? or better than a comprehensive achievement assessment? or better than a review of the students' past history including report cards and grades?" Much has been written about the relationship between IQ and prediction of school achievement. Regarding that relationship, Kaufman (1990 pg 18) reviewed a number of studies and noted, regarding the correlations, "The overall value of .50 is high enough to support the validity of the IQ for the purpose that Binet originally intended it, but low enough to indicate that about 75% of the variance in school achievement is accounted by factors other than IQ." On related matters, an entire issue of the Journal of Learning Disabilities (October 1989) was devoted to a debate about the relevance and usefulness of IQ in the assessment and determination of learning

disabilities. Comments and questions have also been raised about the integrity of IQ scores for learning disabled children. (Communiqué 1988 and 1994). Does the presence of a learning disability affect the scores on an IQ test so as to call into question the issue of current functioning versus potential functioning?

To sum up, let us suggest that by determining the specific qualities measured by a subtest, a factor, or a composite score and by analyzing its relative psychometric integrity, the clinician "may suggest" that a child has a potential for experiencing difficulty in a particular academic area. However, it is important to remember that an intelligence test is not meant to be used as a diagnostic instrument. Rather, a weak performance on a particular subtest or subtests may prove most useful as a compass guiding the course your assessment takes. IQ tests are typically very good instruments for generating a hypothesis about someone's strengths and weaknesses, but they are poor diagnostic instruments for evaluating a learning disability. Particular patterns on an intelligence test may give hints to a possible weakness or disorder, but the assessment of such things is typically done with other tools.

Identifying a person's strengths and weaknesses is a process involving empirical guides while the interpretation of the strengths and weaknesses requires clinical inferences and a broad theoretical base. It may be similar to the issue of sight versus insight. The identification of true strengths or weaknesses gives us some 'sight' but provides little insight. Let's stick to what we see, and not get sucked into becoming soothsayers.

Elliott, C. D., (1990) Differential Ability Scales: Introductory and Technical Handbook, The Psychological Corporation, San Antonio: Tx.

Kaufman, A. S., (1990) Assessing Adolescent and Adult Intelligence. Allyn and Bacon.

Kaufman, A. S., (1979) Intelligent Testing With the WISC-R. John Wiley and Sons: New York.

Kavale, K. A., & Forness, S. R. (1984) A meta-analysis of the validity of Wechsler Scales profiles and recategorizations: Patterns or Parodies? Learning Disability Quarterly, 7, 136-156.

Keith, T. Z., Trivette, P. S., Keith, P. B., & Anderson, E. S. (1992) Profile Analysis with the Wechsler Scales: Patterns, not Parodies. Poster presented at the annual meeting of the National Association of School Psychologists, Washington, D.C.

Lawson, J. S., & Inglis, J. (1987) Micro-interpretation or Misinterpretation? A reply to Forness, Kavale, and Nihira. Learning Disability Quarterly, 3, 253-256.

Sattler, J. M., (1988) Assessment of Children, 3rd ed. San Diego: Jerome Sattler.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792402
HITOMETER



# Importance of Test Norms

When we administer tests to children suspected of having learning disabilities, we must have some concern about who makes up the norming sample of the tests we use. Just because a test is published and known, those facts by themselves do not necessarily mean that the comparison groups were well thought out. On example is given below:

> The original norms for the Halstead-Reitan tests are not well founded. Halstead's "normal" population consisted of 29 subjects (8 women) and 30 sets of scores. Ten of these subjects were servicemen who became available for Halstead's study because the were under care for 'minor' psychiatric disturbances. One was awaiting sentencing for a capital crime (in the state at that time it could have been either life imprisonment or execution. Halstead notes that the subject appeared "anxious"). Four were awaiting lobotomies because of behavior threatening their own life and/or that of others. Two sets of scores were made by one subject, a young man, since he was still waiting at the hospital after two months and so took the test again. This is the group whose test performance defined the unimpaired range for the cutting scores in general use with the Halstead tests. (Bolls 1981)

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792403 HITOMETER

# TEST DESCRIPTIONS

John O. Willis, Ed.D., Rivier College

## Test Description Areas



## Achievement

Kaufman Survey of Early Academic Language Skills (K-SEALS) Kaufman, A. S. & Kaufman N. L. (1993)

> The K-SEALS is an individually administered measure of children's language skills, pre-academic skills and articulation. Both expressive and receptive language skills are assessed. The pre-academic skills evaluated include knowledge of numbers, number concepts, letter and words. The K-SEALS was normed on a national sample of 1,000 children and is intended for children ages 3 years to 6 years, 11 months old.

Kaufman Test of Educational Achievement with Updated Norms (K-TEA NU) Kaufman, A. S. & Kaufman N. L. (1985)

> The K-TEA is an individual achievement test presented on an easel with only one or a few items per page. Items are not multiple-choice. It was normed on a nationwide sample of 2,476 students in grades 1 though 12. Scores can be based on the students age or on the student's grade placement.

Peabody Individual Achievement Test-Revised (PIAT-R) Markwardt, F. C. Jr. (1989)

> The PIAT-R is an individually administered achievement test for children ages 5 years to 18 years, 11 months old, providing assessment in six content areas: General Knowledge, Reading Recognition, Reading Comprehension, Mathematics, Spelling and Written Expression. It was normed on a nationwide sample of 1,563 students in Kindergarten through Grade 12.

Wechsler Individual Achievement Test (WIAT) 1992

The WIAT presents one item at a time without time limits, except for the Written Expression subtest. It offers standard scores, percentile ranks, stanines, and other scores, based either on the students age (four-month intervals through age 12.3, one-year intervals for ages 14 through 19) or the student's grade (fall, winter, and spring norms for each grade), compared to a random, stratified, nationwide sample of 4,252 students of ages 5 through 19 in kindergarten through grade 12. A sample of 1,284 students was given both the WIAT and a Wechsler Intelligence Scale so that students' WIAT Scores could be compared to achievement scores predicted from their intelligence scale scores on the basis of actual test scores from the sample. Achievement scores predicted from intelligence tests fall closer to the mean (Standard score 100, percentile rank 50) than the intelligence scores from which they are predicted.

Wide Range Achievement Test-Third Edition (WRAT-3) Wilkinson, S. S. (1993)

The WRAT-3 is designed to measure reading, spelling, and arithmetic skills in individuals aged 5 to 75, with the possibility of using one of two alternate forms.

(Top of page)

**BEHAVIOR RATING SCALES**

Attention Deficit Disorders Evaluation Scale (ADDES), McCarny, S. B. (1989)

This school version of the scale, used with children ages four to twenty, was designed to provide a measure of Attention Deficit Disorders: inattention, impulsivity and hyperactivity. The standardization sample consisted of 4,876 students ages 4 to 20, from 78 public schools systems in 19 states.

Behavior Assessment System for Children (BASC) Reynolds, C. R. & Kamphaus, R. W. (1992)

The BASC is a multi-method and multi-dimensional approach to evaluating the behavior and self perceptions of children aged 4 to 18 years. The system includes a self-report scale, a rating scale for parents and a rating scale for teachers. It measures numerous aspects of behavior and personality including positive (adaptive) and negative (clinical) dimensions.

Burks Behavior Rating Scales (BBRS), Burks, H. F. (1977)

For use with children grades I through 9, the BBRS are designed to identify patterns of behavior shown by children who have been referred for behavior difficulties at home or in the classroom. It is meant to be a preliminary device for identifying particular problem or patterns of problem a child may be presenting.

Child Behavior Checklist /4-18 (CBLC) and Teacher Rating Form (TRF) Achenbach, T. M. (1991)

The CBCL and TRF are checklists and questionnaires for children ages 2 to 18 years old completed by the student's parent or teacher, describing interests and activities and rating more than 100 potential problems on a 2-1-0 scale. The checklists were normed on a relatively large sample of children with and without known behavioral problems. Percentile Ranks and T scores are based on children without known problem. Competence scales assess reports of school and job perforrr4nce, sports, and social activities.

Conners' Rating Scales, Conners, C. K. (1990)

Brief questionnaires for parents and teachers, focusing on attention, impulsivity, and social problems associated with ADHD. Each item is rated 0 (not at all), 1 (just a little), 2 (pretty much), or 3 (very much). The Comers' Rating Scales are normed for children aged 3 years to 17 years.

Learning Disability Evaluation Scale (LDES), McCarney, S. 8. (1989)

The LDES was designed to be a factor in the determination of the existence of a specific Learning disability. The LDES provides the opportunity to gather performance observations from teachers for students ages 4.5 to 19 years old. The LDES was normed on a total of 1,666 students, from 71 school districts, representing nineteen states.

Vineland Adaptive Behavior Scales (VABS) Sparrow, Balia, & Cicchetti (1984)

The Vineland Adaptive Behavior Scales are not 'tests,' but questionnaires completed by teachers (Classroom Edition) or by an evaluator working with a parent or other care-taker (Survey and Expanded Interview Forms). They include domains of Communication, Daily Living Skills, and Socialization and, for younger students, Motor Skills, with subdomains within each domain. There is also a Maladaptive Behavior scale for the Interview forms. The Interview forms were normed on a representative, national sample of 3,000 persons from birth through age 18. The Classroom Edition was normed on a fairly representative, national sample of 1,984 students from age 3 through 12. Reliability of the scales depends on individual circumstances. The standard scores and percentile ranks are useful.

(Top of page)

**COGNITIVE**

Differential Ability Scales (DAS) Elliott, C. (1990)

The DAS is an individual cognitive abilities test, developed and improved from the British Abilities Scales, for students of ages 2 through 17. It includes verbal, nonverbal (fluid reasoning), nonverbal/spatial, achievement, and special diagnostic tests. The DAS was carefully normed on a stratified, random, national sample of 3,475 students. It is designed to be interpreted by both individual subtests and clusters of subtests, not merely by the total score, which is an important consideration for students with unusual patterns of strengths and weaknesses. Different subtests are used at the lower preschool, upper preschool, and school

age levels. It is handy to be able to compare the achievement tests to the cognitive ability tests within a single instrument, but unfortunately the achievement subtests measure only oral reading of words, written spelling, and math computation.

For an expanded description press here

Kaufman Assessment Battery for Children (K-ABC) Kaufman , A. S. & Kaufman N. L. (1983)

The K-ABC, standardized on a national, stratified sample of 2,000 children, is designed to assess the intelligence and achievement of children ages 2 to 12. The Mental Processing Scales measure the child's ability to solve problems with emphasis on the thinking process used. The Achievement Scale measures acquired knowledge and skills.

Kaufman Brief Intelligence Scale (K-BIT) Kaufman, A. S. & Kaufman N. L. (1990)

The K-BIT, standardized on a national sample of 2,022 people from ages 4 to 90, is designed as a brief individually administered measure of verbal and nonverbal intelligence of people ages 4 through 90. The test takes approximately 15 to 30 minutes to administer and was developed specifically to be used for screening and related purposes.

Test of Nonverbal Intelligence- Second Edition (TONI-2) Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1990)

The TONI-2 is a language-free measure of cognitive ability for individuals ages 5 years old to 85 years, 11 months old that was standardized on a national sample of 2,764 people in the same age range. The TONI-2 offers an administration and response format that eliminates language and reduces motoric and cultural factors. The basis of all of the TONI-2 items is problem solving and the content is abstract/figural.

Wechsler Adult Intelligence Scale-Third Edition (WAIS-III) Wechsler, D. (1998)

The WAIS-R is an individual test that does not require reading or writing, and is intended for adolescents and adults aged 16 years through 89 years old. The Verbal tests are oral questions without time limits except for Arithmetic. The Performance tests are nonverbal problem, all of which are timed and some of which allow bonus points for extra fast work. Test scores and IQ scores are based on the scores of the 2,450 adolescents and adults. Scaled Scores are based on the student's own age group. the WAIS-III allows for the computation of four indexes: Verbal Comprehension, Perceptual Organization, Working Memory, and Processing Speed.

Wechsler Intelligence Scale for Children-Third Edition (WISC-III) Wechsler, D. (1991)

The WISC-III is an individual test that does not require reading or writing, and is intended for children aged 6 years to 16 years, 11 months old. Verbal subtests are oral questions without time limits except for Arithmetic. Performance subtests are nonverbal problems, all of which am timed and some of which allow bonus points for extra fast work. Subtest scores, IQ scores, and factor index scores are based on the scores of the 2,200 children originally tested in a very carefully designed, nationwide

sample.

For an expanded description press here

Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R) Wechsler, D. (1989)

The WPPSI-R is an individual test that does not require reading or writing, and is intended for children aged 3 years through 7 years, 3 months old. The Verbal subtests are oral questions without time limits except for Arithmetic. The Performance subtests am nonverbal problems, all of which are timed and some of which allow bonus points for extra fast work. Subtest scores, IQ scores and factor index scores are based on the scores of the 1,700 children originally tested in a very carefully designed, nationwide sample.

(Top of page)

**PERCEPTION, MEMORY AND VISUAL MOTOR SKILLS**

Bender Gestalt Test of Visual Motor Perception

The Bender asks the student to draw pencil copies of nine fairly complex geometric designs. Erasing is allowed. Some examiners ask the student to attempt re-drawing the designs from memory immediately after copying the last one. Koppitiz's scoring system for children up to age 12 emphasizes the overall shape (Gestalt) of the design and, unlike many similar tests, does not penalize minor errors on details.

Jordan Left-Right Reversal Test, Jordan, B. T. (1990)

This test can be administered either individually or to groups to measure visual reversals in children aged 5 through 12. The standardization sample consisted of 3,000 children aged 5 through 12, administered the test in average classroom settings.

Test of Visual-Perceptual Skills (non-motor) (TVPS), Gardener, M. F. (1988)

The purpose of the TVPS is to determine a child's visual-perceptual strengths and weaknesses based on non-motor visual-perceptual testing. The TVPS was standardized on a group of 962 children, ranging in ages from 4 years through 12 years, 11 months.

(Top of page)

**PROJECTIVE**

Children's Apperception Test

The child is shown a series of pictures and asked to create a story with a beginning, middle and end based on the pictures.

Draw-A-Person

The student is asked to draw a picture of a person. The evaluator often asks the student standardized or individualized questions about the drawing.

House-Tree-Person

The student is asked to draw a picture of a house, of a tree, and of a person. The evaluator often asks the student standardized or individualized questions about the drawings.

Incomplete Sentence Blank

The student completes a series of incomplete sentences. The evaluator often asks the student questions about the responses.

Kinetic Family Drawing

The student is asked to draw a picture a family doing something together. The evaluator often asks the student standardized or individualized questions about the drawing.

Thematic Apperception Test

The individual is shown a series of pictures and asked to create a story with a beginning, middle and end based on the pictures.

(Top of page)

**SPEECH AND LANGUAGE**

Assessing Semantic Skills through Everyday Themes (ASSET) Barrett, M., Zachnwn, L, & Huisingh, R. (1988)

ASSET is a test of receptive and expressive semantics for children ages 3 years through 9 years, 11 months. The ASSET was standardized on 706 school-age children.

Boehm Test of Basic Concepts-Revised (BTBC-R)

The BTBC-R assesses a child's understanding of basic language concepts.

Clinical Evaluation of Language Fundamentals-Revised, (CELF-R) Semil, E., Wiig, E. H., A Secord, W (1987)

The CELF-R is intended as a tool for the identification, diagnosis and follow-up evaluation of language skill deficits for children aged 5 years through 16 years, 11 months.

Expressive One-Word Picture Vocabulary Test-R (EOWPVT-R) Gardner, M. F. (1990)

The purpose of the EOWPVT-R is to obtain an estimate of a child's verbal intelligence by means of the child's acquired one-word expressive picture vocabulary. The EOWPVT-R is intended for children aged years to 11 years, 11 months and was standardized on 1,118 children.

Goldman-Fristoe Test of Articulation (GFTA)

The GFTA measures a child's articulation using basic words at all position: initial, radial and final. It assesses articulation both in single word situations and in simple sentences.

Language Processing Test (LPT) Richard, S. & Hanrier, M. A. (1985)

The LPT assesses a child's ability to attach meaning to language and effectively formulate a response for children aged 5 years through 11 years, 11 months. The LPT was standardized on 497 children from Wisconsin and Florida.

Oral and Written Language Scales (OWLS) Carrow-Woolfolk, E. (1996)

The OWLS are an individually administered assessment of receptive and expressive (oral and written) language for children and young adults. The OWLS are intended for children aged 3 years to 21 years, 11 months on the Listening Comprehension and Oral Expression scales and for children 5 years to 21 years, 11 months on the Written Expression scale. Standardization consisted of a national sample of 1,313 children and young adults.

Peabody Picture Vocabulary Test-Revised (PPVT-R) Dunn, L. M. & Dunn L. M. (1981)

The PPVT-R measures single-word, receptive or listening vocabulary by presenting the student with spoken words and, for each word, showing the student four pictures from which to chose the best match for the word. The test was normed on a large, representative, national sample (4,200 children and 828 adults) and serves its very narrow purpose well.

Receptive One Word Picture Vocabulary Test (ROWPVT) Gardner, M. F. (1985)

The purpose of the ROWPVT is to obtain an estimate of a child's one-word hearing vocabulary based on what he/she has learned from home or formal education. The ROWPVT is intended for children ages 2 years to 11 years, 11 months and was standardized on 1128 children.

Test of Auditory Perceptual Skills-Revised (TAPS-R) Gardener, M. F. (1996)

The primary purpose of the TAPS-R is to assess various areas of a child's auditory-perceptual skills. It is a measure of the child's ability to perceive auditory stimuli and process the stimuli. The test assesses the child's strengths and weaknesses in seven areas of auditory-perceptual skills. The TAPS-R was standardized on a national sample of 1038 children ages 4 years to 12 years, 11 months.

Test of Language Development-Third Edition (TOLD-3) Newwcomer & Hammill (1997)

The TOLD-3 consists of two separate tests: a primary version for children aged 4 years through 8 years, 11 months normed on a national sample of 1000 children; and an intermediate version for children aged 8 years through 12 years, 11 months normed on a national sample of 779 children. Both versions have the same objective: to measure the expressive and receptive competencies in the major components of linguistics. The TOLD-3 can be used to identify children who are significantly below their peers in language proficiency and can determine children's specific strengths and weaknesses in language skills.

Test of Problem Solving (TOPS) Zachnwn, L., Jorgensen, C., Huisingh, R, & Barrett, M. (1984)

The TOPS is an expressive test designed to assess children's thinking and reasoning abilities critical to events of everyday living. The tasks the TOPS assesses include explaining inferences, determining causes, negative why questions, determining solutions, and avoiding problems. The TOPS is intended for students ages 6 years to 11 years, 11 months and was normed on a sample of 456 children.

Test of Written Language-Third Edition (TOWL-3) Hammill and Larsen, 1996

The TOWL-3 includes 'contrived' writing tests and a 'spontaneous' writing sample for which the student writes a 15-minute story about one of two pictures. The TOWL-3 was normed on 2,217 students of ages 7 through 17 tested in 25 states during 1995. The contrived subtests include writing sentences to demonstrate understanding of written vocabulary words, writing from dictation sentences which are scored for spelling and "style" (punctuation and capitalization), rewriting illogical sentences so they make sense, and writing compound and complex sentences to combine simple sentences. The spontaneous subtests are all based on the story that the student writes about a picture. The story is scored for Contextual Conventions (form, punctuation, and spelling), Contextual Language (sentence structure, grammar, vocabulary, and spelling), and Story Construction (prose, action sequencing, and theme).

The Word Test-Revised, Huisingh, R, Barnett, M, Zachman, L., Blogden, C., & Orrmn, J. (1990)

The Word Test - Revised is a diagnostic test of expressive vocabulary and semantics for children ages 7 years to 11 years, 11 months. It is designed to assess a child's ability to recognize and express the critical aspects of his/her lexicon. The word test was standardized on a sample of 805 children.

(Top of page)

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792404 HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# OUTLINE FOR THE EVALUATION OF A TEST

John Willis, Ed.D. & Ron P. Dumont, Ed.D., NCSP

Title of the Test
Author(s)
Publisher
Date of Publication
Date of previous editions
Forms Available
Cost

1. Manual
    - Does a manual accompany the test?
    - Adequacy
    - Is there a separate technical manual and at what cost?
2. Stated purpose of the test?
    - Definition of Construct
    - "Dumbed Down Tests" (tests designed for adults or adolescents redone for children)
3. Does the name of the test reflect the test content?
    - Do the names of the Individual Subtests (where applicable) reflect the content?
4. Form(s) of the items: (Oral, Hands-on, Multiple-choice, Fill-ins, etc.)
    - Are there problems with this form or content?
    - Is scoring ambiguous?
    - Do the items appear to measure what was intended? (e.g., Do reading items really test memory?)
5. Basis of the arrangement of the items in the test?
    - Subtests
    - Scales
    - Spiral Omnibus
    - Random
    - Hierarchical
    - Homogeneity: Changes within subtests
    - Distinctness
    - Sexism and other biases
6. Printing, format and arrangement of test items.
    - Easels and other hardware

- Color use: does it help or hurt?
- Readability

7. Protocols

    - Room to write
    - Answers to examinee
    - Report forms
    - Clarity
    - Ease of use
    - Do they encourage use of confidence bands? Do they offer 90% and 95% bands?

8. Directions for administration
    - Clarity and adequacy?
    - Location (manual/protocol/both)
    - Flexibility
    - Age appropriateness
9. Directions to the examinee
    - Clarity and adequacy
    - Natural or Stilted
    - Boehm's basic concepts
    - Alternative directions
10. Time limits and bonuses?
    - Are they justified?
    - Are there alternatives?
11. Teaching items?
    - Scored or unscored
    - Adequacy of instructions
    - Can you teach over and over?
12. Test materials
    - Child safety
    - Ease of use
    - Durability
13. Scoring
    - Is scoring easy? objective? subjective? arbitrary? agreed upon?
    - Are there adequate samples of correct answers?
    - Rotation errors: differences on tests
    - Are printed norms tables also available?
    - Is computer program necessary?
    - Is computer program provided?
14. Raw scores conversions
    - Interpolation

Which standard scores are reported?
Age scores: Why/why not
Grade scores: Why/why not
Percentiles
Standard scores:
Z
T
Stanines
Deviation quotients (M=100, s.d.=15 or 16)
Others

15. Standardization groups?
    Total
    Number per year of age
    National representation
    Breakdowns
16. For what groups is the test designed?
    Recent
    Relevant
    Representational
    Age
    Grade
    Sex
    SES
    Education
    Geographic regions
    Urban vs. rural
    Ethnicity
    Disabilities
17. Reliability coefficients
    Internal (split halves)
    Alternate forms
    Test retest
    practice effect
    inflation of r
    Length of test
    Test retest interval
    $SE_m$
    $SE_{est}$
    Inter-rater reliability
18. Validity
    For what purpose?

- Content
  - are the questions appropriate ?
  - are there enough questions?
  - level of mastery being measured?
- Criterion
  - concurrent vs. predictive
- Construct
  - Discriminant use vs. divergent use

19. Factor analysis
    - Exploratory
    - Confirmatory
    - Rotations
    - Different groups
    - Variance
    - Common
    - Error
    - Specificity
20. User friendliness
    - Administrator
    - Client: Take it yourself
21. References
    - Antiquity
    - Authors of bibliography
    - Relevance to current edition
22. Interpretation
    - Base rate
    - Definitions for constructs and shared abilities
    - Multiple comparison tables (critical values)
    - Significance vs. abnormality (unusualness vs. importance) (scatter)
    - Testing the metaphysically handicapped (dead)
    - What a difference a day makes
    - Table Games
    - Floor and Ceilings
    - Descriptive terms
    - Errors
    - Cautions

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792405 HITOMETER

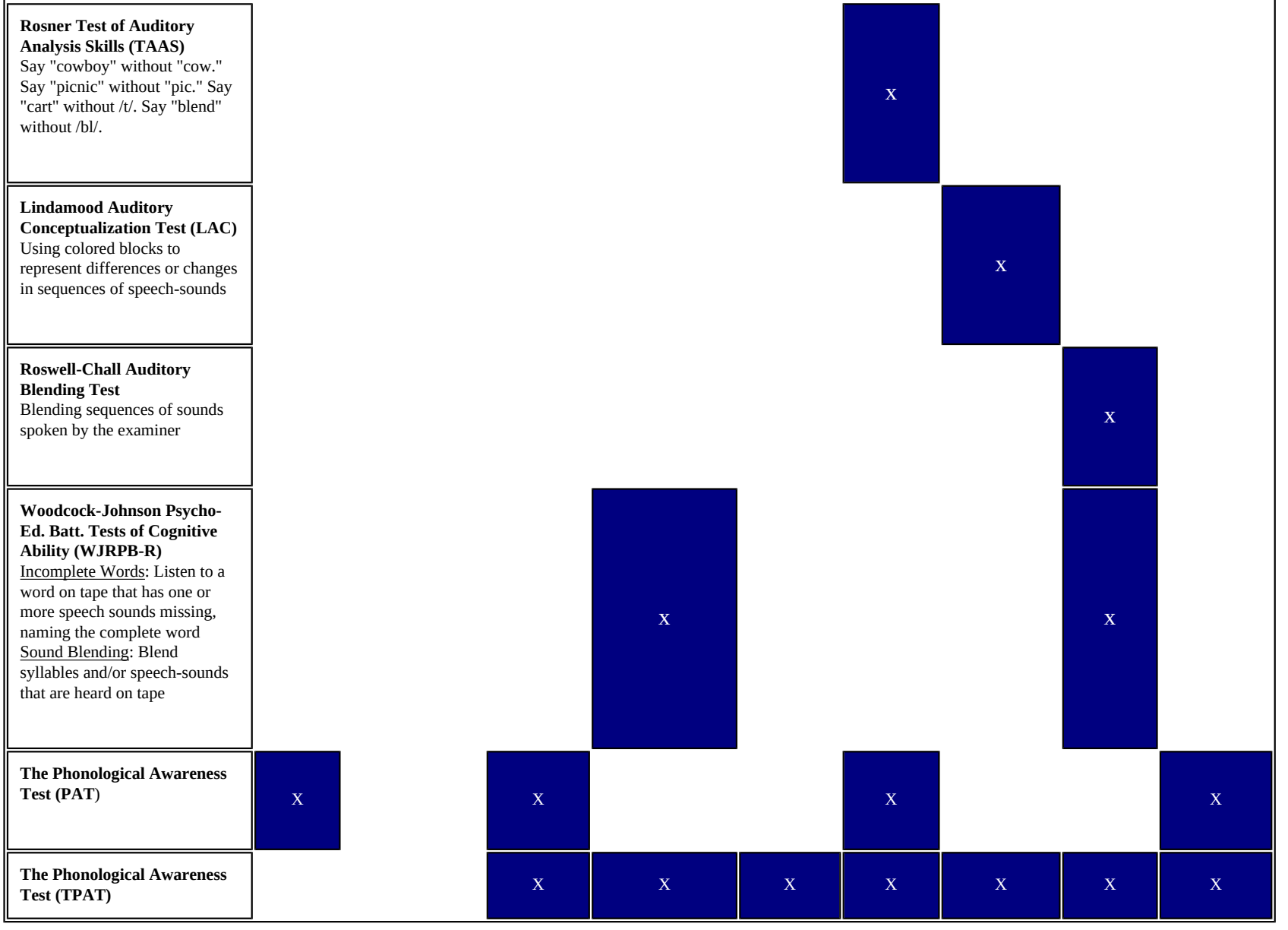
Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# Tests Measuring Aspects of Phonological Awareness

Melissa Farrall, Ph.D. Rivier College

| Test | Rapid Naming | Word Discrimination | Rhyming | Segmentation | Isolation | Deletion | Substitution | Blending | Graphemes |
|---|---|---|---|---|---|---|---|---|---|
| **Test of Auditory-Perceptual Skills-Revised (TAPS-R)** Auditory Word Discrimination subtest Identify whether two words spoken by the examiner are the SAME or DIFFERENT | | x | | | | | | | |
| **Goldman-Fristoe Woodcock Test of Auditory Discrimination (GFWAD)** Listen to words on tape that sound alike and pointing to the matching picture, repeating specified sounds in taped words, reading & spelling nonsense words, choosing pictures matching taped words broken into speech-sounds | | x | | | x | | | x | x |
| **Test of Phonological Awareness (TOPA)** Marking pictures of orally presented words that are distinguished by the same of different sound in the word-final position | | | | | x | | | | |




Phonological awareness refers to a general appreciation of the sounds of speech as distinct from their meaning. When insight includes an understanding that words can be divided into a sequence of phonemes, the finer-grained sensitivity is termed phonemic awareness
See Brody (1994) pp. 388-392 and Rath, L. in Brody pp. 101-140.

RAPID NAMING: speed and fluency of retrieving and saying specific words (a weakness in this skill is often associated with similar difficulties naming letters, retrieving letter sounds, and recognizing written words. There are two subcategories:

Confrontation naming: fluently saying the words represented by pictures
Constrained naming: fluently generating words in categories, e.g., words beginning with B or names of vegetables.

GRAPHEME: The smallest unit in the writing system of a language; letter symbols. (Brody, 1994, p. 390)

BLENDING: integrating separate sounds into a word, e.g., turning /c/ /a/ /t/ into "cat." "Blending refers both to recognizing separate sounds as a word when the sounds are spoken by the examiner and to integrating the separate sounds when "sounding out" (phonetically decoding) a word from print for oneself.

WORD DISCRIMINATION: recognizing how a spoken word differs from another spoken word. Word discrimination is usually tested by dictating two or three words (e.g., "cat, cat" or "cat, cap" or "pack, pack, pack" or "pack. pat, pack") to a student and asking the student if they were the same or different. Testing can be amplified by asking the student to repeat the words, say what the different sounds were, or say where (beginning, middle, or end) the difference was.

ISOLATION:

DELETION: removing a single sound within a word (e.g., "Say 'manhole' without the 'man,'" or "Say 'blend' without the /b/."

ISOLATION: identifying a single sound within a word, e.g., "What is the middle sound in 'bip'?")

RHYMING: Two words rhyme if they end in the same sounds. Rhyming is tested by having the student name words that rhyme with a given word or by asking the student if two or more spoken words rhyme.

SEGMENTATION: breaking words into component sounds, e.g., saying "cat" as /c/ /a/ /t/, or syllables, e.g., saying "segmentation" as "seg-men-ta-tion."

SUBSTITUTION: substituting one sound for another in a word, e.g., "Say 'bat.' Now change the /b/ to an /h/ and say it," or "These three different-colored blocks represent the sounds in 'cat.' I want you to take another block and change it to 'bat,'" or "Read this (dog). Now change the 'd' to an 'h' and read it again."

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792406
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# CTONI Comments

## by John Willis

I like the CTONI with some reservations.

It seems to be very tough for some of the kids with cerebral palsy with whom I have used it. There was some research years ago by Martin Berko that indicated that students with cerebral palsy did better on the Stanford-Binet L-M than on the original Columbia Mental Maturity Scale. Marty hypothesized, as I recall, that the Columbia's demands for categorizing and sequencing made the test more difficult for the brain-injured kids with cerebral palsy. I wonder if the same thing is operative with the CTONI.

A lot of the pictures seem to be outside the experience of many of the children I see who have mobility disabilities. The graphics are a little difficult in some cases.

The ceiling rule is, I find, awfully abrupt. A lot of the kids I see are erratic in their responses and are able to pass a lot of items (with clearly thoughtful decisions, not random guesses) beyond the first set of three errors in five items. I can report that the student would have done "better," "a heck of a lot better," or "wicked better" if we could have counted items beyond the ceiling, but I wish it had been normed with a looser ceiling rule.

I like the fact that you have separate subtests, not just a global scale. You can report a pictorial score vs. a geometric score, which is handy. With some arithmetic, you can even figure out prorated scores for analogies, categories, and sequences. I am mildly suspicious of the norms, which treat a score for 3 subtests, whether pictorial or geometric, identically. I would have guessed that the norming sample would not have had identical performances on the two different sets of items.

Test-retest reliability is reported (on a sample of 63 third and eleventh graders) to be in the .80s for most subtests, high .80s for the pictorial total, and low .90s for the geometric total and test total. Reported correlations with the TONI-3 are modest (.75 and .77), which may be a good thing. Reported correlations for 43 elementary students with LD are .82 with the TONI-2, .74 with the PPVT-R, .76 with the WISC-III VIQ, .70 with PIQ, and .81 with the FSIQ. Unfortunately, the CTONI does not give the means of the tests. With 32 deaf students of ages eight to eighteen, the correlation with the WISC-III PIQ was .90 with subtests ranging from .70 to .90. The highest correlations were with BD and OA.

Also, the CTONI norms are at one-year intervals, so a child of age 6-0 and one of age 6-11 are on the same page in the norms tables. The total norming sample was 2,901 (2,129 ages 6 through 18) tested mostly in 1995. 91% of the sample had no known disability.

**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**

792407
HITOMETER

Dumont-Willis

THE LEADER IN GLOBAL EDUCATION
FAIRLEIGH DICKINSON UNIVERSITY

# ORAL AND WRITTEN LANGUAGE SCALES

## John O. Willis

The Oral and Written Language Scales (OWLS; Carrow-Woolfolk, 1996) includes a Written Language Scale designed to assess writing skills. It is has norms for ages 5-0 through 21-11 and fall and spring norms from the spring of kindergarten through the spring of grade 12 and takes about 20 minutes (10 to 30) to administer. In addition to the total score, the OWLS Written Expression provides percentile ranks for nine special Skills Areas categories in the domains of Conventions, Linguistics, and Content. The child is administered one of four overlapping blocks of items (1-20, 12-24, 21-32, or 25-39) depending on age. However, the use of Rasch-type Ability Scores for the blocks allows out-of-level testing for children with unusually weak or strong writing skills. A maximum of 70 to 82 raw score points is available for each block, permitting considerable bottom and top for each examinee. There is a variety of item types, including copying printed words and sentences; writing letters, words, and sentences from dictation; and writing sentences and paragraphs according to specific, oral instructions. The OWLS Written Expression scale may be administered in small groups "with examinees 8 years and older who are being assessed for reasons other than placement decisions" (Carrow-Woolfolk, 1996, p. 33), as was done in some cases during the standardization.

**Scoring Categories**

The OWLS Written Expression Scale does not have subtests, but does provide reproducible Descriptive Analysis Worksheets permitting calculation of percentile ranks and determination of strengths and weaknesses for nine of the 15 Skills Areas at each year of age.

***Conventions***

- Letter Formation (Item set 1-20): writing legible, recognizable letters
- Spelling (Item sets 1-20, 12-24, 21-32, 25-39): spelling and correct use of words and avoiding omissions
- Capitalization/Punctuation (Item sets 1-20, 12-24, 21-32, 25-39): capitals and terminal punctuation, but only essential internal punctuation
- Conventional Structures (Item sets 12-24, 21-32, 25-39): three items assessing formatting of a letter or paragraph
- Linguistics
- Modifiers (Item sets 1-20, 12-24, 21-32): correct use of adjectives and adverbs
- Phrases (Item sets 1-20, 12-24, 21-32): use of prepositional, infinitive, gerund, and participial phrases
- Question Form (Item sets 1-20, 12-24): correct phrasing (not punctuation) of direct and indirect questions
- Verb Forms/Sentences (Item sets 1-20, 12-24, 21-32, 25-39): construction of sentences and use of correct verb forms

- Complex Sentences (Item sets 1-20, 12-24, 21-32, 25-39): correct construction of complex sentences

***Content***

- Meaningful Content (Item sets 1-20, 12-24, 21-32, 25-39): writing responses that make sense and meet the requirements of the instructions
- Details (Item sets 1-20, 12-24, 21-32, 25-39): two items counting the number of correct details recalled when writing a story that was read to the student
- Coherence (Item sets 12-24, 21-32, 25-39): logical, continuous connection of sentences in a response
- Supporting Ideas (Item sets 12-24, 21-32, 25-39): formulating and expressing ideas in support of an argument
- Word Choice (Item sets 12-24, 21-32, 25-39): precision, vividness, and appropriateness of vocabulary expressing ideas or information
- Unity (Item set 25-39): coherence of the response; focus of all sentences on one idea.

**Scores**

Each item is scored on the basis of one to nine very explicit, completely independent scoring rules in the various Skills Areas so a particular response might, for example, receive points for meaningful content, details, and supporting ideas, but not for spelling and capitalization/ punctuation. Not all scoring criteria require perfection (e.g., "No more than one incorrect word."). Each of the several criteria for each item has a maximum possible score of one to four points. Scoring criteria for each item are listed on the record form and explained, with many examples, in the Manual. Further helpful information is given in the Glossary. The raw score for an item set is converted to a Rasch-type Ability Score with 68%, 90%, or 95% confidence limits, which can be converted to standard scores ($M$ = 100, $SD$ =15), percentile ranks, normal curve equivalents, stanines, and grade- and age-equivalent scores. Age-based norms are provided at three-month intervals for ages 5-0 through 8-11, at four-month intervals for ages 9-0 through 13-11, six-month intervals for ages 14-0 through 18-11, and 12-month intervals for ages 19-0 through 21-11. Grade-based norms are provided at half-year intervals from the spring of kindergarten through the spring of twelfth grade.

**Standardization**

The standardization group contained a representative national sample of 1,373 students stratified to match the U.S. census data for 1991 on the basis of age, sex, and four categories each of mother's education, race/ethnicity, and geographic region. An additional 185 students with language impairments, mental handicaps, learning disabilities, hearing losses, and reading delays were tested for clinical validity studies. There were 115 to 124 children at each year of age from 5 through 11, and average of 65 per year from 12 to 15, and an average of 42 per year from 16 through 21. There were 111 to 159 children per grade from kindergarten through grade 6 and an average of 51 children per grade for grades 7 through 12.

**Reliability**

Internal consistency reliabilities range from .77 to .94 (*Mdn* $r_{xx}$ = .87). Test-retest reliabilities over 18- to 165-day intervals for a sample of 84 examinees were .66 with a mean gain of 0.0 standard score points for ages 8 through 10 and .83 with a mean loss

of 1.3 points for ages 16 to 18. [Corrected for the variability of the norm group (*SD* = 15), the reliabilities were .88 and .87, respectively.] Interrater reliability for four raters scoring 60 protocols ranged from .91 to .98 at various ages (*Mdn* $r_{xx}$ = .94).

**Validity**

Construct validity of the OWLS Written Expression Scale is based on extensive development efforts to match the content and format of the test to language theory (e.g., Carrow-Woolfolk, 1988, 1996; Carrow-Woolfolk and Lynch, 1981). The correlations with other writing tests reported in the Manual were .67 with the Kaufman Test of Educational Achievement (K-TEA) Comprehensive Form Spelling (n = 31), .78 with the Peabody Individual Achievement Test-Revised (PIAT-R) Spelling (n = 31), and .77 with the PIAT-R Written Language Composite (n = 31). [Corrected for the variability of the norm group, these were .82, .73, and .71, respectively.] Correlations with total reading scores were .75 with the K-TEA, .84 with the PIAT-R, and .80 with the WRMT-R (n = 29). [Corrected for the variability of the norm group, these were .86, .80, and .87, respectively.]

Correlations with oral language tests were .57 with the PPVT-R (n = 100), .74 with the Clinical Evaluation of Language Fundamentals-Revised (CELF-R), [Corrected for the variability of the norm group, these were .62 and .79, respectively.] .57 with the OWLS Listening Comprehension Scale (n=1,364), .66 with the OWLS Oral Expression, and .67 with the OWLS Oral Composite.

Correlations with intelligence tests included .61 with the Wechsler Intelligence Scale for Children, 3rd ed. (WISC-III) Verbal IQ, .51 with the Performance (nonverbal) IQ, and .59 with the Full Scale IQ (n = 34). [Corrected for the variability of the norm group, these were .72, .64 and .70, respectively.]. Correlations with the Kaufman Brief Intelligence Test (K-BIT) were .62 with Vocabulary, .32 with Matrices, and .52 with the Composite (n = 62). [Corrected for the variability of the norm group, these were .67, . 41, and .58, respectively.]

Students in the clinical samples with language impairments, mental handicaps, and learning disabilities (reading) and learning disabilities (undifferentiated) groups all scored significantly lower than students in matched control groups. Students in a Chapter One Reading Program scored lower (*M* = 90.7, *SD* = 11.7) than the norming sample.

**Comment on the OWLS Written Expression Scale**

The OWLS Written Expression Scale offers a brief, but comprehensive assessment of writing skills. The use of multiple, direct and indirect items rather than a single writing sample does not penalize students who write only a few words when given fifteen or twenty minutes to write a story or a letter to a friend. For students who are willing and able to write a longer sample, you can supplement the OWLS with one of the various story- or letter-writing subtests or with an informal writing sample, e.g., "You have been asked to give a graduation speech at a teachers' college. Your speech will be the last advice these college graduates will hear before they go out to teach students like you for the next 40 years. Please give them the most important advice you think they need. You may use this paper to write an outline, web, or notes before you begin."

The 241-page Manual is clear, explicit, and helpful. After a little practice, scoring and interpretation quickly become efficient.

The OWLS Written Expression Scale can be purchased and used alone, but it can also be used with the OWLS Listening Comprehension and Oral Expression Scales, which are brief, but very useful instruments a psychologist can use to assess the impact of oral language difficulties on the student's functioning and to decide whether to make a referral for an in-depth language assessment. The OWLS Manuals include data on statistical significance and base rates of differences among the three scales.

---


Carrow-Woolfolk, E. (1988). *Theory, assessment and intervention in language disorders: An integrative approach.* Philadelphia: Grune & Stratton.

Carrow-Woolfolk, E. (1996). *Oral and Written Language Scales: Written Expression Scale Manual.* Circle Pines, MN: American Guidance Service.

Carrow-Woolfolk, E., & Lynch, J. I. (1981). *An integrative approach to language disorders in children*. San Antonio, TX: The Psychological Corporation.



**Content on these pages is copyrighted by Dumont/Willis © (2001) unless otherwise noted.**