RIVIER UNIVERSITY

**DIVISION OF EDUCATION**

# **SPECIALIST IN THE ASSESSMENT OF INTELLECTUAL FUNCTIONING PROGRAM**

AND

**ASSOCIATION OF SPECIALISTS IN ASSESSMENT OF**

**INTELLECTUAL FUNCTIONING (ASAIF)**

[**http://www.asaif.net**](http://www.asaif.net)

**Comments on Reports 3/8/13 # 245**

If you wish to receive copies of this newsletter, email me at [johnzerowillis@yahoo.com](mailto:johnzerowillis@yahoo.com). Email versions include notices of ASAIF and other conferences and jobs in and near New Hampshire.

The **Association of Specialists in Assessment of Intellectual Functioning (ASAIF)** sponsors educational activities supporting the assessment of intellectual functioning, including this newsletter, co-sponsored by the Specialist in Assessment of Intellectual Functioning program at Rivier University,[[1]](#footnote-1) evening dinner-and-training events called "Shorties," and workshops. **ASAIF is now authorized by NASP to offer CPD credits.**  **If you have topics on which you would like ASAIF to do a workshop or Shorty, please tell me at** [johnzerowillis@yahoo.com](mailto:johnzerowillis@yahoo.com)**. We have worked with school districts to co-sponsor workshops in the districts. We are happy to travel outside New Hampshire if someone wants to pay the speaker's travel expenses.**

**CONTENT**

**Processing Speed (G*s*) and RAN Tests** almost always use relatively easy (technical term: "no brainer") items because they are assessing speed. [Correct Decision/Reaction Time/Speed or G*t* measures may be more challenging.] High accuracy is the expectation for Processing Speed and RAN tests. Low accuracy is the exception. If we mention in our reports accuracy on Gs and RAN tests ["Ralph worked slowly, but very accurately on the Coding A subtest (scaled score 1, percentile rank 0.1) . . ."] the reader is likely to focus on Ralph's accuracy on these very easy items (of which he completed, at age 7:8, no more than 15 items of the easy items in two minutes) rather than on his dramatic slowness, which is what the subtest is intended to measure. It is safer, unless accuracy is poor, to mention only the speed. [Yes, it would be really cool if publishers included at least base-rate norms for accuracy on such tests, so we could recognize the rare occasions when we should mention accuracy.] Low accuracy (or significant slowness) on a RAN test should make us wonder how well the examinee knows the items being tested. For example, the problem might not (yet) be slow naming of letters; it might be not knowing some or many of them at all. Assessing speed of naming letters would need to wait until we have taught the letters.

**Score classification labels.** Harping on an old issue that constantly recrudesces, I feel compelled by recent reading to reiterate the necessity of explaining the arbitrary names assigned to test scores by authors and publishers.

Salli-An's intellectual ability on the RIAS was fairly strong, only "moderately below average" (standard score 70), but her Oral Reading Index on the GORT-5 was "poor" (standard score 79), and her phonological awareness was absolutely "below average" (standard score 89) on the CTOPP, so she is clearly reading far below her intellectual ability level.

Moncton scored Average (standard score 85) on the Wechsler Individual Achievement Test (WIAT-III) Reading Comprehension and Average (standard score 109) on the Wechsler Intelligence Scale for Children (WISC-IV) FSIQ, so he clearly needs no further special instruction in reading. We are thrilled with Moncton's progress.

We have two choices. We can use the different, even conflicting, labels provided in the various test manuals (although, as far as I can tell, none at all are offered for Wechsler IQ test scaled scores in the Wechsler manuals), or we can select (or create) one classification scheme (I like stanines, but nobody else does; the Woodcock-Johnson system has the virtue of symmetry, but I don't care for the term "superior") and apply it to all tests. Both choices will inevitably cause confusion. In either case, we simply must use footnotes and appendices to tell and then remind readers what we are doing. This issue was discussed at painful length in Report Comments #240 with a long attachment. If you would like a copy of that attachment and a short PowerPoint (recently edited by Dr. Carol Evans) on test scores, please email me at [jwillis@rivier.edu](mailto:jwillis@rivier.edu) or [johnzerowillis@yahoo.com](mailto:johnzerowillis@yahoo.com) and tell me you want them. If I don't send them, email me again.

Ten Top Problems with Normed Achievement Tests for Young Children. Ron Dumont, Fairleigh Dickinson University, and John O. Willis, Rivier University. [This is an old issue (look at the dates of the references!) and I have shared this little paper with almost everyone I have ever met, but the issue keeps coming up (well, actually, not since late last night), so I repeat it here.]

1. There is no agreed-upon preschool or kindergarten curriculum at national, state, and even, in some cases, local levels. It is difficult to sample a curriculum that does not exist. For higher grades, there is at least some commonality among the many curricula at a given grade level. The same skill may be placed at very different levels. See, for example, <http://alpha.fdu.edu/~dumont/psychology/WR_vs_LWI.htm>, <http://alpha.fdu.edu/~dumont/psychology/NO_VS_CALC.htm>, and <http://alpha.fdu.edu/~dumont/psychology/MR_VS_AP.htm>

2. Young children are often inconsistent in their responses, which would argue for more items to increase reliability.

3. But young children have short attention spans and they fatigue easily, which requires fewer items.

4. Sampling works well for a large domain. For example, if a child is expected to have a reading vocabulary of 3,000 words, it is pretty easy to estimate reading skill with a 25-word test. However, if a child is expected to have a reading vocabulary of 10 words, your 25-word test could, by pure chance, easily sample all 10 or none of them, giving an inflated or depressed estimate of the child's reading ability. Similarly, many achievement tests for young children have only a few letter-naming items, rather than 52. If a child knows ten to sixteen letters, a ten-item test could easily hit or miss all of them by pure chance. If you test a child on Monday, and the teacher teaches the vowels on Tuesday, that could be the difference between a score of zero and a score of five on the ten-item test.

5. Young children develop new skills so rapidly that norms tables should be divided by weeks, not three, four, or six months. The difference between age 10-0-0 to 10-6-29 and age 10-7-0 to 10-11-29 may be trivial, but the difference between 4-0-0 to 4-6-29 and 4-7-0 to 4-11-29 is tremendous.

6. Item format matters a lot more for younger children. Most ten-year-olds don't care whether an addition problem is presented horizontally or vertically, but five-year-olds do. The space between lines on writing paper and the presence or absence of a dotted midline can be a deal-breaker for most kindergarten students.

7. Item gradients are necessarily very steep for younger children. For example, there aren't any clearly defined steps between not being able to write the letter M and being able to write it.

8. Norming samples are also a huge problem at the preschool level. If you carefully sample geographic regions, parental education and income, and other germane variables, you can be fairly safe in using whatever public and private schools (in the right proportions) are available. However, at the preschool level, there is a huge difference between the Mary Poppins School of Unfettered Self-Expression and Free Play and the John Stuart Mill Preschool of Relentless Academic Pressure. Severely disadvantaged kids from the JSMPRAP Prep are likely to score higher on academic achievement tests than rich kids from the MPSUSEFP. A truly representative national sample (especially with only 100 to 300 kids per year of age) is virtually unobtainable.

9. There often is insufficient floor for young children on achievement tests. See, for example, <http://alpha.fdu.edu/~dumont/psychology/McGee.htm>.

10. Consequently, formal and informal, criterion-based tests (with exhaustive sampling, e.g., all 53 letters, all sums up to ten, etc.); curriculum-based measurement; classroom observations; and work samples are likely to be much more informative than normed tests up to at least a mid-second-grade level of achievement (regardless of actual grade placement).

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* (2), 155-166.

Bracken, B. A., & Walker, K. C. (1997). The utility of intelligence tests for preschool children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 484-502). New York: Guilford Press.

Schultz, M. K. (1988). A comparison of Standard Scores for commonly used tests of early reading. *Communiqué,* Newsletter of the National Association of School Psychologists, *17* (4), 13.

**Raw Scores of Zero.** Ron Dumont, Fairleigh Dickinson University, & John Willis, Rivier University. [This is another old issue on which I have often harped, but it also continues to crop up like stink bugs, so here I go again.]

We are severely allergic to zero raw scores. Jeri J. Goldman (1989) is the pioneer in the study of zero raw scores. We have tried to contribute to this vital field with the Evaluation of Sam McGee at <http://alpha.fdu.edu/~dumont/psychology/McGee.htm>.

A zero raw score can sometimes mean that the student lacks an essential skill that is not the intended target of the test. The WJ III test of Editing is a good example, used by the publisher, Riverside, in communications with WJ III purchasers. Editing is intended to measure a specific editing skill in kids who can read. If the student cannot read (especially if it is because nobody has tried to teach this four-year-old how to read yet), then what should have been a test of proofreading skill inadvertently becomes just another reading test.

More important, though, is that zero raw scores near the bottom end of a test's age range can produce insanely high scores (for example with Goldman's and our dead students).

Another problem is that you don't know whether it was a high zero or a low one. Say, for example, the zero raw score yields a standard score of 65. Was the student almost capable of getting one item correct and leaping to a standard score of 78? Or, if the test had had sufficient bottom, would the student still have gotten a zero, reflecting a true ability at a standard-score level of 20?

Sampling error becomes a horrendous problem with zero raw scores. We should worry when scores are based on only a few correct responses, much less zero. Suppose a beginning first grader can read 25 words at sight and has no other word attack skills. By sheer chance, that kid's set of 25 words might include 10 words on the WJ-III, 5 on the KTEA-II, and none on the WIAT-III (or any other combination). The closer you are to the bottom of the test, the more sampling error will bite you. Schultz's (1988) comparison of standard scores for commonly used tests of early reading, discussed above, found that different achievement tests yielded widely different scores for the same test performance, e.g., standard score of 98 on one test and 65 on another. The tests have changed since 1988, but the issue has not.

A zero raw score might mean that the student simply missed the whole point of the exercise, even though the student was reasonably competent in the underlying ability that the test was intended to measure. A particular child might, for example, be able to demonstrate verbal abilities pretty well on an analogies test, but not on a similarities one. Daniel (1986) found, for another example, that block-design tests with flat tiles and with three-dimensional cubes measured the same construct in sixth grade children. However, the 3-D cubes might confuse some younger children who could have demonstrated their abilities with flat tiles (Elliott, 1990, p. 48).

Even if a zero raw score should happen, by accident, to produce an accurate standard score, you certainly would not have much data to work with. We strongly recommend that if you get within about one SEm of a zero raw score, it would be prudent to find some other measure of that ability that had more bottom.

Daniel, M. H. (1986, April). *Construct validity of two-dimensional and three-dimensional block design*. Paper presented at the annual convention of the National Association of School Psychologists, Hollywood, FL.

Elliott, C. D. (1990). *Differential Ability Scales introductory and technical handbook*. San Antonio: The Psychological Corporation.

Goldman, J. J. (1989). On the robustness of psychological test instrumentation: Psychological  
evaluation of the dead. In Glenn G. Ellenbogen (Ed.) *The Primal Whimper: More Readings from the Journal of Polymorphous Perversity.* New York: Ballantine, Stonesong Press.

Schultz, M. K. (1988). A comparison of Standard Scores for commonly used tests of early reading. *Communiqué,* Newsletter of the National Association of School Psychologists, *17* (4), 13.

**Pet Peeve of the Week.**  What are "support services"? Do they include direct instruction, are they just reteaching as needed? Are they merely hand-holding? They really should be defined when they are cited in reports (not to mention in Individualized Education Programs).

**Significance Levels and Base Rates.** I have been getting a lot of questions about this issue, so I created an example that one person has found useful. In the table below, I have put a red **X** in each row to show where each difference between a score and the examinee's own mean (AB -2.0, FG +2.4, MN -1.6, PQ +2.0) would fall.

The first step is to determine **in advance** what you consider a significant difference. For example, p < .05 means that, if such a large difference would occur simply by random variation only 5 times in 100, you are willing to abandon the idea that it is likely to be random variation and entertain the possibility that the observed difference represents a real and reliable difference. If you are a cautious skeptic, you might hold out for p < .01, requiring the difference to be so large that it would occur randomly no more than 1 time in 100. If you tend to cross streets from between parked cars without looking, you might accept p < .10 or even p < .15. In this example (and in my own testing), I am going with p < .05. In my test manuals, I often use a highlighter to mark the .05 columns in the tables.

The second step is to determine **in advance** what you consider an uncommon difference. Significant differences are not necessarily uncommon. Human abilities vary. Nobody is too surprised, for example, that Sammy is getting an A in Math but a C+ in Spanish or that Pat is a great golfer, but can't pole vault. An uncommon difference is one that is so large that no more than xx percent of people taking the test when it was normed showed such large differences. You have to decide for yourself what percent you would consider uncommon. I am going with 10 percent. It is a popular, but not universal, choice. Again, I attack my test manuals with a highlighter. Following are our fake data.

AB (–2.0) is not significant and not uncommon. The difference would have to be 2.05 or more to be significant and 2.67 or more to be uncommon.

FG (+2.4) is significant: 2.4 is at least as big as 1.91. It is also uncommon. The difference must be 2.33 or bigger to be uncommon. The difference of 2.4 is bigger than 2.33.

MN (–1.6) is not significant and not uncommon. MN would have to be ±2.22 or bigger to be significant and ±3.50 or bigger to be uncommon.

PQ (+2.0) is significant (at least as big as 1.92), but not uncommon (not as big as 3.00).

In all your manuals, just highlight the .05 significance and 10% base rate columns (or your personal choices).

Positive and negative values for differences (scores higher or lower than the mean) obviously matter for interpretation, but significance and base rate are based on absolute values: +2.3 is the same as –2.3.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Significance Level | | Base Rate | | | | | |  |
| Sub-test | Difference  from  Mean | .15 | .05 | 1% | 2% | 5% | 10% | 15% | 25% |  |
| AB | – 2.0 | 1.67 | 2.05 | 4.17 | 3.67 | 3.00 | 2.67 | 2.33 **X** 1.67 | |  |
| ***not significant*** | | | | ***base rate between 15% and 25% - not uncommon*** | | | | | |  |
| FG | + 2.4 | 1.56 | **1.91** | 3.67 | 3.33 | 2.67 | **X** 2.33 | 2.00 | 1.67 |  |
| ***significant!*** | | | | ***base rate between 5% and 10% - uncommon!*** | | | | | |  |
| MN | – 1.6 | 1.92 | 2.22 | 5.50 | 4.80 | 4.10 | 3.50 | 3.10 | 2.40 | **X** |
| ***not significant*** | | | | ***base rate more than 25% - not uncommon at all*** | | | | | |  |
| PQ | + 2.0 | 1.57 | 1.92 | 4.67 | 4.33 | 3.67 | 3.00 | 2.33 | **X**2.00 |  |
| ***significant!*** | | | | ***base rate 25% - not uncommon*** | | | | | |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Significance Level | | Base Rate | | | | | |  |
| Sub-test | Difference  from  Mean | .15 | .05 | 1% | 2% | 5% | 10% | 15% | 25% |  |
| AB | **– 2.0** | 1.67 | 2.05 | 4.17 | 3.67 | 3.00 | 2.67 | 2.33 **X** 1.67 | |  |
| FG | **+ 2.4** | 1.56 | **1.91** | 3.67 | 3.33 | 2.67 | **X** 2.33 | 2.00 | 1.67 |  |
| MN | **– 1.6** | 1.92 | 2.22 | 5.50 | 4.80 | 4.10 | 3.50 | 3.10 | 2.40 | **X** |
| PQ | **+ 2.0** | 1.57 | 1.92 | 4.67 | 4.33 | 3.67 | 3.00 | 2.33 | **X**2.00 |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | **Significance Level** | | **Base Rate** | |
| Sub-test | Difference  from Mean | Not Significant | Significant  .05 | Uncommon 10% | Not Uncommon |
| AB | score – mean | anything less than 2.05, such as +2.04 or –2.04 | 2.05 | 2.67 | anything less than 2.67, such as +2.66 or –2.66 |

If a difference is not significant, it is not a difference. It should not be called a difference in the report. There is no such thing as a so-called "non-significant difference." It don't mean a thing.[[2]](#footnote-2) **TANSTAAND**[[3]](#footnote-3)

A significant difference that is not uncommon can still be important. More than 10% of the WISC-IV norming sample (12%) had FSIQs of 82 or lower, so a FSIQ of 82 would not be uncommon, but it nonetheless might help explain some of a student's school difficulties.

Some test manuals run their tables of base rates in the opposite direction just to see if we are on our toes.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | Significance Level | | Base Rate | | | | | |
| Sub-test | Difference  from  Mean | .15 | .05 | 25% | 15% | 10% | 5% | 2% | 1% |
| AB | **– 2.0** | 1.67 | 2.05 | 1.67 **X** 2.33 | | 2.67 | 3.00 | 3.67 | 4.17 |
| FG | **+ 2.4** | 1.56 | **1.91** | 1.67 | 2.00 | 2.33**X** | 2.67 | 3.33 | 3.67 |
| MN | **– 1.6** | 1.92 | 2.22 | **X** 2.40 | 3.10 | 3.50 | 4.10 | 4.80 | 5.50 |
| PQ | **+ 2.0** | 1.57 | 1.92 | 2.00**X** | 2.33 | 3.00 | 3.67 | 4.33 | 4.67 |

I do not recommend cluttering up our reports and the minds of our readers with detailed discussions of these statistics. The first time I mention a significant difference, I insert a footnote that goes something like this: "Even the very best tests cannot be perfectly reliable. Lucky and unlucky guesses, good and bad days, or barely beating or going over a time limit, for example, can make scores vary randomly. Therefore, differences between scores also vary by chance. In this report a 'significant difference' is one that is too great to occur by random variation more than 5 times in 100." The first time I mention an uncommon difference, I insert a footnote along these lines: "Even significant differences are not necessarily uncommon. Human abilities vary. In this report, an 'uncommon difference' is a significant difference so large that it was found in no more than 10 percent of the persons taking the test when it was normed." After using each footnote once (sometimes in the same sentence, e.g., The difference between Ralph's Above Average score on Picture Disambiguation and the average of all of his scores on the FUBAR [Average for his age] was significant[[4]](#footnote-4) and uncommon[[5]](#footnote-5)." Once I have used my footnotes once each, I can just write "significant" and "uncommon" without a care in the world and without further annoying my readers. I certainly do not need to explain each uncommon difference with a long, confusing, parenthetical discussion about how this "difference was seen in only 3% of the children in the norming sample for the FUBAR" and that "difference was seen in only 1% of the children in the norming sample of the FUBAR." Not only does that comment get old really quickly, but also remember that we were required to select in advance one significance level (e.g., p < .05) and one base rate level for uncommonness (e.g., ≤ 10%) and stick with them. Picking or changing our criteria after the fact is like firing at a blank piece of paper and then drawing a bull's eye around the hole.

**STYLE**

**Phonology.** Sadly most people who are not educators (and some who are ☹) do not know much about phonology. It is probably prudent to mention in our reports that phonology involves sounds in spoken words, while phonics involves the connection between those sounds and written letters. Phonology can be done in the dark; phonics cannot (except with Braille). We could even observe that it is, therefore, kinda tough to learn phonics without some skill in phonology. I find it is also helpful to offer examples of the bizarre stunts demanded of students on phonology tests. We could, for example, explain, ". . . elision (*blend* without the /b/ sound would be "lend" and *blend* without the /l/ sound would be "bend") . . ." or ". . . phoneme reversal (the sounds in *cat* would make "tack" in reverse order, and the sounds in *knife* would be "fine" in reverse order. . ." or ". . . blending (hearing the sounds /k/…ă…/t/ and recognizing "cat") . . ."

**Between and from.** What's up with, for example, "between 2003 to 2005" or "a score between 90 to 109"? I am seeing and hearing that strange construction increasingly often. We don't write that "Manchester is between Nashua to Concord." It is "between Nashua and Concord."

Similarly, a written phrase such as "from 2003-2005" leaves me wondering, "from 2003-2005 to what?"

**Agreement of subject and verb.** For some reason, I am seeing many reports that use singular subjects and plural verbs, for example, "The test of oral vocabulary, semantic categorization, verbal abstract reasoning, and meaningful categorization were Malthusia's lowest." Yes, it were.

**BOOKS AND JOURNALS**

Wheelan, Charles. *Naked Statistics: Stripping the Dread from Data.* (New York, NY: W. W. Norton, 2013. ISBN 978-0-393-07195-5). Wheelan teaches public policy and economics here in New Hampshire at Dartmouth. I am tempted to take a class from him. This book is a very clear, non-technical, humorous explanation of basic statistical concepts with minimal use of math. All of us reading this newsletter are, of course, comfortable with math and advanced statistics, but you might get some helpful ideas for explaining statistics in your reports and team meetings. I am really enjoying this book.

John O. Willis

Senior Lecturer in Assessment, Rivier University

Assessment Specialist,

Regional Services and Education Center

419 Sand Hill Road

Peterborough, NH 03458-1616

(603) 924-0993 [johnzerowillis@yahoo.com](mailto:johnzerowillis@yahoo.com)

1. Neither ASAIF nor Rivier University is in any way, shape, or form responsible for the quirky opinions in this newsletter. They cannot be blamed for what is written here. [↑](#footnote-ref-1)
2. <http://en.wikipedia.org/wiki/It_Don't_Mean_a_Thing_(If_It_Ain't_Got_That_Swing)> [↑](#footnote-ref-2)
3. <http://en.wikipedia.org/wiki/There_ain't_no_such_thing_as_a_free_lunch> [↑](#footnote-ref-3)
4. Even the very best tests cannot be perfectly reliable. Lucky and unlucky guesses, good and bad days, or barely beating or going over a time limit, for example, can make scores vary randomly. Therefore, differences between scores also vary by chance. In this report a "significant difference" is one that is too great to occur by random variation more than 5 times in 100. [↑](#footnote-ref-4)
5. Even significant differences are not necessarily uncommon. Human abilities vary. In this report, an 'uncommon difference' is a significant difference so large that it was found in no more than 10 percent of the persons taking the test when it was normed. [↑](#footnote-ref-5)