The following article is the opinion of the author. It is has not been peer reviewed by the Canadian Psychological Association (CPA) nor is it the official position of CPA. It is intended to inform and to stimulate discussion. Other opinions are welcome.

The WISC–IV: Observations and Cautions for those used to the WISC-III

Mary Ann Evans, Clinical Psychology: Applied Developmental Emphasis Program
University of Guelph, Guelph, Ontario

The year 2003 was an unprecedented one in seeing the release within the short span of a few months of revised versions of two of the most commonly used cognitive batteries for assessing children. While the Stanford-Binet Intelligence Scale – 5th edition is a departure from its previous 4th version in incorporating features of its still earlier forms, the Wechsler Intelligence Scale for Children-IV seems to have under gone a "tinkering", and with that there are several pitfalls. It is similar enough to the WISC-IV that on the surface old WISC-III habits would appear to serve one well, and the instrument might seem to be an improved version. However close inspection shows that the WISC-IV is not without its glitches, and that aside from the obvious differences between the third and fourth editions in the subtests, there are subtle differences between the two that can lead to an inappropriate or questionable transfer of habits from the older to newer edition. Moreover flaws in the WISC-IV can lead to questionable application of the numbers it generates.

I have listed my observations for readers in the hopes that my experience in teaching the WISC-IV in the winter semester of 2004 will be of value to new users until an updated version of Jerome Sattler's text, A*ssessment of Children; Cognitive Application*, or similar helpful volume, or an improved WISC-IV appears as a beacon of light. These observations fall under four general categories: **Differences between the WISC-III and WISC-IV** (subtests, scales, scale compositions, response formats, scoring, time, etc.); the **FSIQ and Four Indexes** (VIQ and PIQ are not among them); **Substituting One Subtest for Another** (or rather not substituting one subtest for another); **Slips and Errors on Their Part** (e.g., someone was asleep when labelling the bottom grid on the normal curve on the back of the protocol); and **Slips and Errors on the Part of Examiners** (to err is human but, alas in this context, unprofessional and potentially harmful to clients).

DIFFERENCES BETWEEN WISC-III AND WISC-IV

The core and supplementary subtests are different, with six WISC-III subtests either deleted or designated as supplementary on the WISC IV and six new subtests added.

- The WISC-IV no longer has Picture Arrangement, Object Assembly and Mazes. Arithmetic, Picture Completion, and Information have been changed to supplementary subtests.

- Added are Letter-Number Sequencing; Matrix Reasoning and Picture Concepts as core subtests, and Word Reasoning, and Cancellation as supplementary subtests.

- Core subtests retained from the WISC-III are Block Design, Similarities, Vocabulary, Comprehension and Coding.

- Previously supplementary subtests that are now core subtests are Digit Span and Symbol Search.

- There are new items and format changes in many of the WISC-IV versions of these subtests.

Thus over half of the test is different.

The WISC-IV is more of an actuarial and less of a clinical instrument than the WISC-III.

- The technical manual explains that the removal of Pictures Arrangement, Object Assembly and Mazes was done to make way for new subtests and to reduce the emphasis on performance under time constraints. The deletion of Picture Arrangement and Object Assembly also makes the WISC-IV easier to give in that these subtests had required both skilful timing and skilful manipulation of the test materials. These subtest with relatively few but longer manipulatory items were replaced with multiple - choice/point-to-the-option subtests where one simply adds up the correct choices.

- The biggest effect of the deletion of these two subtests along with Mazes is that there are fewer materials for the child to manipulate. As a result, the clinician is less able to observe the child's approach to problem solving and response style, or to isolate what part of a problem was missed or created particular difficulty. Thankfully Block Design has been retained as core. However it is now the first subtest. This is not a wise choice given that children frequently can tell that they have been stumped or unsuccessful on a design and may quickly lose confidence. (Previously Picture Completion was first on the WISC-III and typically enjoyed by children who could happily guess at an answer even when unsure.)

In summary, the WISC-IV is less engaging for children, has fewer materials for the child to manipulate, and yields less qualitative information to interpret the numbers.

There are fewer subtests where speed of performance is a factor.

- The WISC-III included four subtests with time bonuses: Arithmetic, Picture Arrangement, Object Assembly, and Block Design starting with item #4. In addition three subtests had a time limit for item completion: Picture Completion (20 second limit), and Coding (120 sec) and Symbol Search (120 sec). On the WISC-IV, Arithmetic, while having time limits for each item, has no time bonus, and Block Design now provides times bonuses starting with item 9. Coding, Symbol Search and Picture Completion have the same time limits as on the WISC-III, but as noted above Picture Completion is now only a supplementary subtest.

Although speed of performance does not affect scores, clinicians should take care to note time to completion to distinguish between children who can solve a problem quickly, versus within the time limit, versus outside the time limit as much as opportunity allows. This one of the few bits of qualitative information you will be able to gather.

The change is what are core versus supplementary subtests has altered the composition of the Full Scale IQ.

- In the WISC-III only 2 of the 10 core subtests (Coding and Digit Span) required less "acumen", higher order reasoning, or flexibility of thought. Research has also shown that these subtests are also the ones that children with intellectual challenges typically perform relatively better on within their test profiles and gifted children relatively less well. (See technical manual for statistics to confirm this). In the WISC-IV, 4 of the 10 core subtests (Coding, Digit Symbol, Letter-Number, Symbol Search) are of this nature.

- While 2% of the population will still fall in "Very Superior" and "Extremely Low" classifications of general intellectual ability, it may not be the same composition of individuals within these groups.

- In addition the increased emphasis on these subtests may mean that children with learning disabilities, for whom these subtests are typically relatively weakly performances (see technical manual for statistics on children with reading disabilities) may obtain lower Full Scale IQs.

While data appears in the technical manual as to the difference between FSIQ on the WISC-III and WISC-IV for an unselected sample ages 6-16, no data on this appears for special groups. Research is needed to determine whether the new composition of core subtests on the WISC - IV differentially affects the full scale IQ children in categories of cognitive exceptionality.

The WISC-IV is longer.

- 10-30 seconds instead of 15-20 seconds are suggested as sufficient for a child to begin to respond.

- Five subtests from the WISC-III now have more items: 14 versus 12 for Block Design, 8 versus 7 for Digit Span, 21 versus 18 for Comprehension, 33 versus 30 for Information, and 34 versus 24 for Arithmetic. While not all items are given, if ceiling is not reached there are more items to give than on the WISC-III.

- The WISC-III manual stated that 50-70 was required to administer the core subtests. The WISC-IV manual states that 65-80 minutes are needed. In fact only 50% of the children in the standardization sample completed the core subtests with 67 minutes. This is not the reduced testing time that the manual claims.

Thus it is not only that the less engaging format of the WISC-IV makes it seem longer when giving it, it is longer.

The 3 IQ scores have been reduced to one.

- Gone is the VIQ and PIQ. (See next section)

THE FSIQ AND FOUR INDEXES (Or "What's in a Name? Would a rose by any other just as sweetly smell?")

The 10 core (or 15 core and supplementary subtests) are clustered on four indexes. Core subtests Block Design, Picture Concepts, Matrix Reasoning form the Perceptual Reasoning Index; Similarities, Vocabulary, and Comprehension the Verbal Comprehension Index; Digit Span and Letter-Number Sequencing the Working Memory Index; and Coding and Symbol Search the Processing Speed Index. All form the Full Scale IQ.

Verbal Comprehension Index (Vocabulary, Comprehension, Similarities)

- According to the manual, the VCI can be substituted for the VIQ in clinical decision-making and other situations where the VIQ was previously used or required. Indeed, clinicians have rightly or wrongly relied on the VIQ and PIQ in determining eligibility for programs, funding, exceptionality designations, and diagnoses. Given that the manual reports a correlation is .87 between the VCI of the WISC-IV and the VIQ of the WISC-III, the VCI can be regarded as comparable to the VIQ. Similarly given that the three subtests of the VCI on the WISC-IV are the same as three of the four subtests of the WISC-III VCI, clinical guidelines developed for the WISC-III for interpreting the VCI remain applicable for the WISC-IV.

Perceptual Reasoning Index (Block Design, Matrix Reasoning, Picture Concepts)

- According to the manual, Matrix Reasoning and Picture Concepts were designed to measure fluid reasoning and the change in name from Perceptual Orientation Index on the WISC-III to the Perceptual Reasoning Index (PRI) on the WISC –IV "reflects increased emphasis on fluid reasoning abilities in this index". In addition Object Assembly was dropped "to get a purer reasoning index", and Picture Completion was dropped as it is "more a measure of visual discrimination, attention to visual detail and perceptual ability". This leads one to ask why the Index score was not renamed Fluid Reasoning Index.

- Moreover despite these changes and the accompanying rationale, the manual contends that one can substitute the PRI for the PIQ in clinical decision-making and other situations where the PIQ was previously used or required. However the correlation between the PRI and PIQ according to the manual is .74, meaning that the two share only 55% variance (versus 76% for the WISC-III VIQ and WISC-IV VCI.) In short one should be cautious about using the PRI in the same way as the POI, given the different nature of the core subtests within them and correlation between the two. While the emphasis on fluid reasoning is probably a good thing, the PRI may not be as useful, for example, in the diagnosis of a nonverbal learning disability as the PRI was on the WISC-III.

Working Memory Index (Letter-Number sequencing; Digit Span)

- Letter Number Sequencing is one of two subtests forming Working Memory Index. The first nine items can be repeated verbatim for full points and therefore tap short term, not working memory.

- Digit Span is the other subtest on this index. As in the WISC-III, the first half consists of digits to be repeated in the same order, a measure of short-term memory.

- Thus this index is no more a measure of Working Memory than it's precursor on the WISC-III was a measure of Freedom from Distractibility.

- The WISC-IV does provide norm-referenced scores separately for digits forward and digits backwards in Table A 8, allowing the examiner to get a purer measure of working memory. However given that the reliability of Digits Backwards is less than .80 at ages 7, 8, 9, and 10, examiners should not rely on this or the discrepancy analysis of digits forwards versus backwards to draw conclusions.

Processing Speed (Coding and Symbol Search)

- The measurement of processing speed is presented in the manual as something that is especially important to assess in children. However how one goes about this is limited and muddied. Among other things, Coding taps fine motor co-ordination, incidental learning, working memory, and a willingness to forgo neatness for speed. Symbol Search requires spatial skills in differentiating fine details and mirror images. The two subtests correlate just .53 with each other.

- Cancellation, "developed to provide a supplementary PS subtest", correlates just .40 and .32 with the other Processing Speed subtests (see table below regarding substituting one subtests for another.) Moreover it is a very poor measure of *g* (just 7% of the variance) leading one to wonder why was it included in a scale of intelligence. (All other subtests have at least 26% of their variance attributable to *g)*. The reliability of Cancellation is also lowest of all the subtests.

And then there's *g...* (for which there is no index but there is still a FSIQ to reflect it)

- Good measures of *g* (at least 50% of variance attributed to *g*) are Vocabulary, Information, Similarities, Arithmetic, Word Reasoning, and Comprehension.

- Fair measures of *g* (26-49% of variance attributed to *g*) are Block Design, Matrix Reasoning, Picture Completion, Letter-Number Sequencing, Symbol Search, Picture Concepts, Digit Span, and Coding.

- A poor measure of *g* (only 7% of variance attributed to *g*) is Cancellation.

SUBSTITUTING ONE SUBTEST FOR ANOTHER

According to the manual can substitute one subtest per index to a maximum of substitutions per FS IQ. A major problem is that this allows for many different possible combinations of subtest scores and leads one to question the need for rigorous adherence to standardization procedures. These substitutions may be merited in extraordinary circumstances, but the correlational statistics provided in Table 5.1 of the technical manual (summarized below) should lead one to be very cautions in this practice. The bottom line? Substitute rarely if at all.

Note on this table: a = correlation of core subtest to corresponding index without that subtest
b = correlation if core subtest to corresponding index including that subtest

		Core Subtests		
		Vocabulary	Similarities	Comprehension
Verbal Comprehension Index with *r* =	a	.79	.74	.70
	b	.91	.89	.86
Information with *r* =		.75	.70	.62
Word Reasoning with *r* =		.66	.62	.58
Information with VCI *r* = .77				
Word Reasoning with VCI *r* = .70				

Comment: Information correlates as highly with the core subtests and the VCI as the core subtests intercorrelate and correlate with the VCI. This is less so for Word Reasoning.

		Core Subtests		
		Block Design	Matrix Reason.	Picture Conc
Perceptual Reasoning Index with *r* =	a	.56	.61	.50
	b	.81	.84	.77
Picture Completion with PRI *r* = .57		.54	.46	.39
Picture Completion with VCI *r* = .55				

Comment: Picture Completion correlates less well with the core subtests than they intercorrelate and correlates equally well with the VCI as the PRI.

		Core Subtests	
		Digit Span	Let-Num Sequencing
Working Memory Index with *r* =	a	.49	.49
	b	.86	.86
Arithmetic with WMI *r* = .57		.47	.51
Arithmetic with VCI *r* =.63!			
Arithmetic with PRI *r* =.62!			

Comment: Arithmetic correlates less well with the core subtests than they intercorrelate and correlates more highly with the VCI and PRI than with the WMI.

		Core Subtests	
		Coding	Symbol Search
Processing Speed Index with *r* =	a	.53	.53
	b	.88	.87
Cancellation with PSI *r* = . 41		.40	.32
Cancellation with FSIQ *r* = .26			

Comment: Cancellation correlates less well with the core subtests than they intercorrelate, and correlates poorly with the FSIQ.

SLIPS AND ERRORS ON THEIR PART

- Similarities now has age-differentiated starting points. However, clinicians may use judgement and start at a lower point. Note the that this may have consequences: If child fails the first two items at an age entry point beyond item 1, and the examiner then proceeds backwards to establish basal, except for the one sample item there is no teaching the concept of unless one gets back to item #2. In contrast if child starts at item 1 to begin with, there are two items plus the sample for which a correct response is modelled by the examiner to help the child establish the response set. Given that there are only 4 items before the highest age-differentiated starting point, it is safer to give them all rather than skip them.

- For many subtests the instructions are that “Children suspected of intellectual deficiency should start with item 1”. This is a misphrasing. It really means “start with the sample and then item 1”.

- For Digit Span the instructions are to “Administer both trials of each item, even if child passes trial 1”. This is a misphrasing and really means “Administer both trials of each item, regardless of performance on trial 1", because the second trial is to be administered if the child fails or passes trail 1.

- For Comprehension there is faulty use of “And” and “or” in the generals scoring criteria for some of the items. This can be seen by comparing general criteria for 2-point responses with specific sample responses that should be given scores of 2. See for example, items 2 (“and” should be “or”), 3 (“or” should be “and”), and 15 (“or” should be “and”) if the sample responses are anything to go by. They are supposed to be, because the manual advises examiners to “first attempt to match the child’s response to the sample responses”. (p. 40).

- Substituting Letter-Number Sequencing with Arithmetic: If the child cannot count to “X” as one of the two qualifying trials of Letter-Number Sequencing (this subtest being one of two subtests on the Working Memory Index), then the examiner is to discontinue and administer Arithmetic. The first items of Arithmetic also require the child to count to “X” and beyond “X” in the next two items. Therefore the child who does not pass the qualifying trial on Letter-Number Sequencing likely will score 0 on Arithmetic. Despite the advice in the manual, common sense dictates that this “0” reflects an inability to count, tells us nothing of working memory, and hence cannot be a substitute for L-N Sequencing on the Working Memory Index.

- Matrix Reasoning: What’s the answer to item 26? I surveyed several professors and graduate students on this, only half of whom chose the correct answer. There are 10 more items after this for ages 12 to 16.

- Word Reasoning: What’s the answer to item 23? There are many possibilities that are not noted in the sample responses. And why the procedure of presenting one clue, then the same clue and another, then the same two and a third? This is irritating to the children, there is no differential scoring according to how many clues are required to get the correct answer (nor should there be), and it extends the testing time.

- Process-Level Discrepancy Analysis of Block Design with and without time bonuses. Examiners are asked to compare Block Design scores with and without time bonus points as part of the Process-Level Discrepancy Comparisons to arrive at more detailed information about a child’s performance. However, time bonuses are available only after item 8. If the child reaches ceiling before item 9 or completes only a small number of

items for which a time bonus can be gained there is no (or there is an inadequate) sampling of performance with and without time bonuses. There is no warning of this in the manual before examiners consult Table B.9 to compare the scores. The computerized scoring software also announces that there is no significant difference between the two, even though there may never have been any opportunity to find one!

- Process Analysis Discrepancy of Digit Span and Block Design. Table B.9 is not broken down by age so that for children at any age between 6 and 16, the same magnitude of discrepancy between Digits Forward and Digits Backwards is required for statistical significance. However the reliability of Digits Forward ranges from .79 to .85 at the different age bands, and the reliability of Digits Backwards ranges from .68 to .86. A particularly bad combination is at age 7 where the reliabilities are .79 and .69 respectively. Moreover on both Digits Forwards and Digits Backwards (particularly the latter), a difference on 1 raw score can make for a 2- standard scores point change. This is not a reliable way of determining differences between auditory short-term and working memory.

- I never ask for a computer-generated report but curiosity caught this cat. As a result I know that there is at least one glitch in WISC-IV writer. The report generated this: "Linda performed much better on abstract concept formation and categorical reasoning tasks *that did not require verbal expression (Similarities = 13)* than on abstract concept formation and categorical reasoning tasks *that required verbal expression (Picture Concepts = 9)".* Two paragraphs later it continues " Linda achieved her lowest score on the Picture Concepts subtest…This subtest is designed to measure fluid reasoning and abstract categorical reasoning. It invokes verbal *concepts but does not require verbal responses (Picture Concepts scaled score = 9*. (Italics mine.) In other words, the report said Picture Concepts required verbal expression in one paragraph and did not require verbal expression in another paragraph. I don't know why anyone would want to become brain dead, let alone admit to a lack of skill by using computer generated interpretive reports. Moreover the quote from the computer-generated report shows that this software cannot be trusted and should never be used to generate ideas (let alone be comprehended by the average reader.)

- On the back page of the record form there is an error in normal curve graph. The bottom line is labelled Percentile Rank but details standard scores. This hardly inspires confidence in the care taken vis a vis the rest of the test.

SLIPS ON THE PART OF EXAMINERS (to err is human but, alas in this context, unprofessional and potentially harmful).

- **Negative transfer from the WISC-III.** In addition to attending to minor procedural modifications**,** examiners should take care not to rely on the habits they have developed in scoring the WISC-III because some of the items reused on the WISC-IV have different scoring for the same responses noted in the WISC-III. For example, for item 13 of Comprehension "a lie" used to receive 1 point. On the WISC-IV it receives 0 points.

- **Clerical errors.** On the WISC-III, considering only the core subtests there were 10 subtest raw scores to convert, four index scores to find, and three IQ scores to derive. The WISC-IV has dropped the VIQ and PIQ, making for 2 less conversions but added an "Analysis Page" with 33 scores one can calculate and look up in table, choosing between one of two tables (e.g., the table for the overall sample or for the ability group), a significance level of .15 or .05 within a table, and of course, the table /column row for

the child's age group. This exponentially increases the chances of making clerical errors, and users do!

- **Errors encouraged by confusions in the record form.** The record form contributes to errors by being confusing in using the term Scaled Score on the Analysis Page rather than the term Composite Score (used on the first page of the record from) to refer to Index Scaled Scores. Examiners using the tables in the manual must be careful in using/entering the appropriate score in calculating discrepancies: Subtest scaled scores are used for comparing Digits Forward and Digits backwards; Composite Scaled Scores, not sums of scaled scores, are used for Comparing Indexes. A second way that the record form contributes to errors in the discrepancy comparisons is that the examiner checks a box to indicate whether the "Basis of Comparison" in Index scores is the "overall sample" or a given "ability level". This differentiation is for the *base rate* of the discrepancy found in Table B2, not the *critical value* for the discrepancy found in Table B1. In Table B1, examiners are to locate the appropriate age-based critical value, and not use the critical value for "all ages" appearing on the last line. Users have mistaken inferred that checking the "overall sample" as the basis for comparison means using the age-collapsed statistics in Table B1.

- It is strongly advised that users of the WISC-IV buy the software to check their calculations and conversions, but ignore the option for generating an interpretive report for reasons outlined earlier. However, using the scoring software is not a fail proof method for avoiding errors. Users still must pay close attention to what significance level, reference group, and comparison base they want, and specify this accordingly when entering raw scores.

- Note that confidence bands in Table A2-6 of examiner manual and in the scoring software are based on Standard Error of Estimate values that take into account regression to the mean. If one is wanting to state the confidence band for the true score on the WISC-IV at the time it is administered it, one should use (according to Sattler, 2001 among others) the Standard Error of Measurement. Although the confidence bands based on these two statistics won't likely be different in the average range of IQ Scores, Index Scores, and subtest Scaled Scores, they will become increasingly different towards the extremes of the distribution. Examiners who are used to looking up the latter confidence bands in Sattler's text or will have to get out their calculators and consult Table 4.33 of the more obscure technical manual. For example, multiply the Standard Error of Measurement by 1.65 for a 90% confidence value; or by 1.44 for 85% confidence value. Note that reports normally contain percentiles; so confidence bands in standard scores then need to be converted to a band in percentile units. This provides further opportunities for clerical errors.

- Think twice before using (and interpreting) the WISC-IV with a six or seven year old for which profile analysis will be important. As noted earlier, it is not an engaging test for young children. Moreover reliability is relatively low for Coding, the separate Forward and Backwards scores for Digit Span, and the separate Random and Structured Scores for Cancellation. What looks like a relative weakness may well be an unreliable score.

- Finally, at this point in time, there are no Canadian norms for the WISC-IV. If one believes in doing ability-achievement discrepancy analyses (and I am not sure that one should be wedded to this), it is not appropriate to use the achievement-ability discrepancy tables in the Wechsler Individual Achievement Test - II manual by entering scores from the American norms of the WIAT-II and American norms on the WISC-IV. One reason is that the table is based on the statistical properties of the WISC-III, not the WISC-IV. A second reason is that standard scores from the WIAT-II using American

norms are substantially higher than is the case using the recently released Canadian norms, making children who indeed are below their peers look as though they are well within the average range. Until this is resolved, and because of the short-comings of the WISC-IV noted in this paper, users may not want to discard their WISC-III kits quite yet.

Mary Ann Evans is a professor in Clinical Psychology at the university of Guelph and has regularly taught its clinical cognitive assessment course over the past 23 years. Comments will be received by evans@psy.uoguelph.ca.