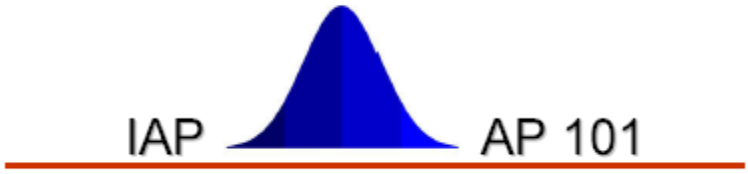


IAP Applied Psychometric 101 Brief # 5:

The Wechsler-like IQ subtest scaled score metric: The potential for misuse, misinterpretation and impact on critical life decisions

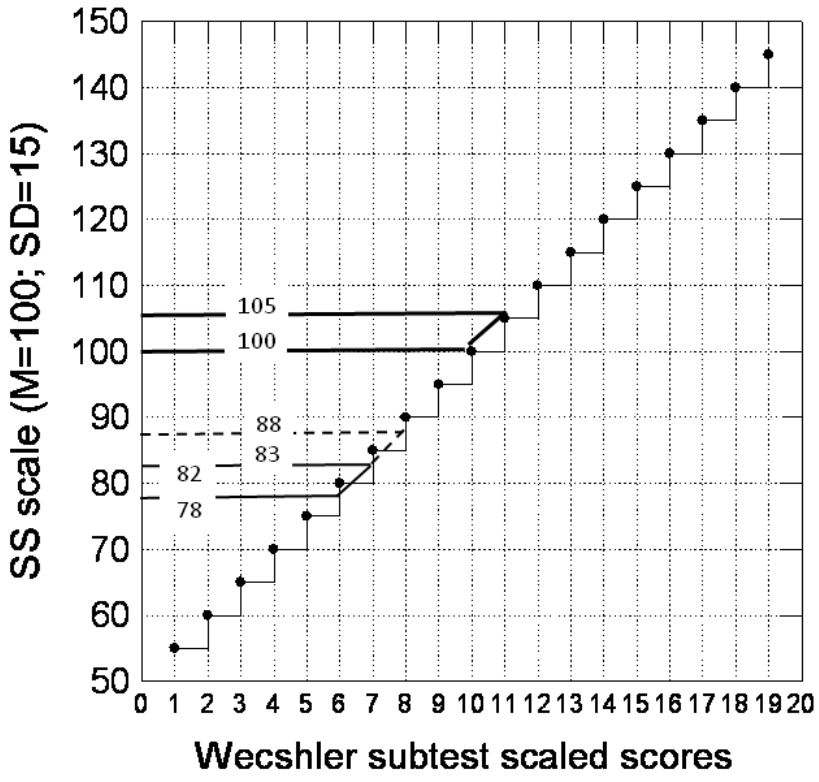
Kevin S. McGrew

I've recently been skimming James Flynn's new book (What is Intelligence: Beyond the Flynn Effect) to better understand the methodology and interpretation of the Flynn effect. Of particular interest to me (as an applied measurement person) is his analysis of the individual subtest scores from the various Wechsler scales across time. As most psychologists know, Wechsler subtest scaled scores (*ss*) are on a scale with a mean (M) = 10 and a standard deviation (SD) = 3. The subtest *ss* range from 1 to 19. In Appendix 1 of his book, Flynn states "*it is customary to score subtests on a scale in which the SD is 3, as opposed to IQ scores which are scaled with SD set at 15. To convert to IQ, just multiply subtest gains by five, as was done to get the IQ gains in the last column."* At first glance, this statement makes the transformation of subtest *ss* to IQ *SS* appear as an easy ("*just* multiply…."; emphasis added by me) and mathematically acceptable procedure without problems. However, on close inspection this transformation has the potential to introduce unknown sources of error into the precision of the transformed *SS* scores. It is the goal of this brief technical post to explain the issues involved when making this *ss-to- IQ SS* conversion.

The *ss 1-19* scale has a long history in the Wechsler batteries. For sample, in Appendix 1 of *Measurement of Adult Intelligence* (Wechsler, 1944), Wechsler described the steps used to translate subtest raw scores to the new *ss* metric. The Wechsler batteries have continued this tradition in each new revision, although the methodology and procedures to calculate the *ss 1-19 values* have become more sophisticated over time. Although the methods used to develop the Wechsler *ss 1-19* scale may have become more sophisticated, the resultant underlying scale for each subtest has not…scores still range from 1-19 (M=10; SD=3). Also, the most recent Stanford-Binet—5th Edition (SB5; Roid, 2003) and Kaufman Assessment Battery for Children-2nd Edition (KABC-II) have both adopted the same *ss 1-19* scale for their respective individual subtests.

Why is this relatively *crude* (to be defined below) scale metric still used in some intelligence batteries when other contemporary intelligence batteries provide subtest scale metrics with finer measurement resolution? For example, the DAS-II (Elliott, 2007) places individual test scores on the *T*-scale (M=50; SD=10), with scores that range from 10-90. The WJ III (McGrew & Woodcock, 2001) places all test and composite scores on the standard score (*SS*) metric associated with full scale and composite scores (M=100; SD=15). The critical question to be asked is "**are there advantages or disadvantages to retaining the historical *ss 1-19* scale or, are their real advantages to having individual test scales with finer measurement resolution (DAS-II; WJ III)?"**

What do I mean by *crude* scales? I asked myself this question as I was reading Flynn's analysis of specific Wechsler subtest score changes over time (i.e., the *Flynn Effect*). As described in his book, Flynn converted the Wechsler subtest *ss 1-19* values (M=10; SD=3) to the *SS* metric (M=100; SD=15). Upon close examination, the *ss/SS* transformation raises a number of issues and questions. These are best demonstrated in the following figure.



The graph plots the traditional Wechsler-like *ss 1-19* (X-axis) and the corresponding *SS* value (Y-axis) for each *ss* from 1-19. For example, in the figure a *ss* of 10 (X-axis) corresponds to an *SS* of 100 (Y-axis). A Wechsler (or SB5 or KABC-II) *ss* of 11 corresponds to an SS of 105. As can be seen in the figure, the *ss/SS* relationship is represented by a ***step function***. That is, for every one point *ss* change there is a corresponding 5 point change in IQ *SS*. Stated differently, each 1 point *ss* increment results in a jump of 5 *SS* points. **The *ss/SS* relationship is not a linear 1-1 function.**

I've represented this 1-to-5 conversion ratio by the two dark lines that map the conversion of *ss* values of 10 and 11 to *SS* values of 100 and 105. It is obvious from the figure that the subtest *ss* is a much less precise measure of ability than the *SS* scale. In particular, the 1/5 *ss/SS* ratio indicates that the *ss 1-19* metric is *20% less precise or sensitive* than the *SS* metric.

Using the information in the figure, one wonders how a specific *ss* score should be interpreted. Although a mathematical transformation provides a specific IQ *SS* for each *ss* (e.g., *ss*10 = *SS*100), in reality each *ss* is best conceptualized as representing a 5 point range of *SS* values. This 5 point range of *SS* scores extends halfway down the *SS* scale towards the *SS* associated with the next lowest *ss* and halfway up the *SS* scale towards the *SS* associated with the next highest *ss*. The lowest pair of solid lines (in the figure) provides more precise guidance. If one assumes that an individual obtains a *ss* = 6 on an individual subtest, this score covers a range of 5 *SS* points. A *ss* = 6 does not represent the single *SS* value of 80, but rather the **range of *SS* values from 78 to 82**. This five point bracket around each *ss* is consistent for all *ss* values from 1-19.

To explain further, I use the analogy of using a tape measure to measure human height. Let's assume we have a simplified measurement tape that retains the 12 in = 1 foot relationship (with no finer gradations between the inch tick marks). This will be our height *SS* scale. If one then crafts a 20 % less precise (more crude) measuring tape, tick marks would not be every inch but would be every 2.4 inches. Using this less precise *ss* tape measure, we measure an individual, who is exactly 68 inches tall (5 feet, 8 inches). Using the *SS* tape measure we would measure this person's height accurately as 68 inches. Conversely, when using the *ss* tape measure the top of the person's head would fall somewhere between the 28th (67.2 inches) and 29th (69.6) tick marks. Given that the person's real height (68 inches) is closer to the 28th tick mark, we would conclude that this person is 67.2 inches tall…and not 68 inches. Since the *SS/ss* metric ratio is 1/2.4, the 28th tic mark would record a value of 67.2 inches for all individuals who have a real *SS* height between 66.0 (-1.2) and 68.4 (+1.2) inches. This is similar to the observation that each single intelligence subtest *ss* (when on the *ss 1-19 scale*; M=10; SD=3) measures a range of 5 *SS* points when individual *ss 1-19* scores are converted to the more precise *SS* scale.

This raises interesting questions when researchers (such as Dr. Flynn) convert subtest *ss* values to the *SS* metric and complete statistical analyses, or calculate group means and SD's based on the *ss 1-19* scale and then convert the summary statistics to the *SS* scale (M=100; SD=15). Is it

possible that the statistical analysis of the new *SS* scores introduces an unknown degree of error or imprecision in the statistical analysis and conclusions? Thinking as a statistician, one can make the reasonable assumption that over a large number of cases (such as in a large standardization sample), *over-and under-estimation* of individual subjects *converted ss-to-SS scores* should occur equally in both directions and thus, cancels out the error due to imprecision of the *ss-to SS* conversion. But can the same be assumed for small research samples (e.g., *n*=100 to 200) where subjects have been administered (in a counterbalanced order) the older and newer (revised) version of an established IQ battery to investigate the Flynn Effect? I'll let quantoids with greater statistical expertise chew on this issue.

However, **the implications are more apparent (and troublesome) in the case of individual (e.g., clinical, forensic, etc.) evaluations**. For example, what does it mean for an individual to obtain a *ss* = 6 when this score represents values from 78 to 82 on the *SS* scale? If an examiner attempts the *ss-to-SS* conversion to allow comparison of individual test scores with scores from other intelligence or achievement batteries that use a finer mental measurement tape measure (e.g., DAS-II; WJ III), is there potential for serious errors in interpretation given that the *ss* = 6 can only be interpreted to **represent a range of scores** from 78 to 82? Furthermore, what if the examiner makes a simple scoring error that produces a 1 raw score change on a subtest which, in turn, changes the persons recorded *ss* value from 6 to 7? It is important to note that in many of the raw score-to-*ss* norm tables provided (for intelligence test batteries that use the *ss 1-19* scale) 1 raw score point change often produces a change of 1 *ss*.

Following the logic and analysis demonstrated for the relationship between *ss* = 6 (which corresponds to a *SS* range of 78 to 82), the *ss* = 7 represents a *SS* range from 83 to 88 (as represented in the figure by the distance bounded by the dashed line and the solid line immediately beneath the dashed line). We are now faced with a situation where the range from the lowest possible *SS* (represented by the "correct" *ss*=6; **78** is the lower end of the *SS* bracket) to the highest possible SS (represented by the "incorrect" *ss* = 7; **88** is the upper end of the *ss* = 7 *SS* bracket) is a **range of 10 SS points**! Of course, this is the most extreme example. A review of the figure shows that it is also equally possible that the swing in range of possible *SS* values (for *ss* = 6 and 7) may only vary from 82 to 83. However, we simply do not how large a potential swing in measured ability is represented by a change of one raw score point when each single *ss* value from 1-19 represents a range of 5 *SS* points each.

Although this is not problematic for a single score here and there, what happens in cases where potentially multiple scoring errors occur across different subtests? For example, let's assume an examiner makes a sufficient number of scoring errors that results in an upward (incorrect) shift to the next *ss* (1 *ss* point increase) on three subtests. Using the WISC-III norm tables for this example,[i] let's assume a person obtained a *sum of scaled scores* (across the complete set of subtests used to calculate the Full Scale IQ) of X that corresponds to an IQ SS of 68 (using Table A.2 in Appendix A of the WISC-III manual---this table is used for all scores reported in this example). Since the three minor errors increased three subtest *ss* values by 1 point each, the

incorrect (but recorded and reported) *sum of scaled scores* is now X+3, which translates to an IQ of 69 or 70 (70 being the cut point most recognized in classification and diagnostic systems for mild mental retardation). What about an individual who should have a sum of scaled scores of Y that converts to a WISC-III IQ SS of 75? Given the same scoring error scenario, the person now receives an incorrect sum of scaled scores of Y+3. This value converts to a WISC-III IQ SS of 76 or 77--which is one or two points above the highest score typically considered as acceptable for a diagnosis of MR/ID in many diagnostic/classification contexts (e.g., eligibility for special education; eligibility for SSA; diagnosis of MR/ID for death penalty cases). And, we have yet to introduce the standard error of measurement (SEM) confidence band (typical rule-of-thumb is ± 5 IQ *SS* points) in the interpretation of these hypothetical scores.

Although we can (and should) argue about the appropriateness of rigid adherence to specific cut-scores when making a diagnosis of MR/ID, a reading of the majority of Atkins court decisions reveals that the courts either set a *bright line cutoff score of 70* or consider the standard error of measurement (SEM) of + or -5 points (which allow scores as high as 75). Clinical judgment is often not allowed in court proceedings or in other guideline-driven eligibility decisions (e.g., SSA and special education classification). Unfortunately, prescribed specific cut-score eligibility criteria often seem to be the norm.

Let's now take the scoring error scenario one step further. Let's assume an examiner made scoring errors (in the upward direction) on enough subtests to increase the total *sum of scaled scores* by 5 points. An individual who had an original (correct) ***sum of scaled scores*** of X (IQ = 68) will now have a *sum of scaled scores* of X+5 (IQ SS = 70 or 71), the second value which crosses the bright line cutoff score of 70. Furthermore, for individuals with a *sum of scaled score*s value of Z (which will produce WISC-III IQ SS of 75--the upper limit allowable for mild MR diagnosis in many contexts), the recorded (but incorrect) *sum of scaled scores* (Z+5) produces an IQ SS of 77 or 78.

How likely are these scenarios? The extant literature (see Ramos, Alfonso & Schermerhorn, 2009 for recent summary—summary table from Ramos et al. is included below—see original article for more readable copy) has shown an unfortunate degree of scoring and administration errors by both novice and experienced psychological examiners on almost all intelligence tests (not just the Wechsler's). For example, **the range of average errors per test record reported in the Ramos et al. table ranges from approximately 11 to 38 errors! Errors in scoring have been reported to produce full scale IQ scores that can differ between examiners as much as 11 IQ *SS* points (Ryan & Schnakenberg-Ott, 2003) and under certain conditions a range of 32 IQ *SS* points (Hopwood & Richard, 2005)!** Thus, these hypothetical scenarios are likely to occur far more frequently during individual clinical and forensic intellectual assessments than we would hope.

From Ramos, Alfonso & Schermerhorn (2009)

Table 1
Summary of Results of Previous Studies Investigating Examiner Errors on the Various Wechsler Scales

Study	Sample	Instrument Investigated	Major Findings
Sherrets, Gard, & Langner (1979)	39 psychologists, interns, practicum students, school psychologists, and psychometricians	WISC	89% of examiners made at least one error; most common errors were in addition of scores
Slate & Chick (1989)	14 graduate students	WISC-R	All subtests were found to have some error; errors on 66% of the protocols resulted in changes to FSIQ
Slate & Jones (1990)	26 graduate students	WISC-R	An average of 11.3 errors per protocol; frequent errors included failure to record examinee responses, incorrect point assignment, and inappropriate questioning
Slate, Jones, Coulter, & Covert (1992)	9 certified psychological examiners	WISC-R	An average of 38.4 errors per protocol, including failure to record responses; errors on 81% of the protocols resulted in changes to FSIQ
Alfonso, Johnson Patinella, & Rader (1998)	15 graduate students	WISC-III	An average of 7.8 errors per protocol; frequent errors included failure to query, failure to record responses verbatim, reporting incorrect FSIQs, reporting incorrect VIQs, and incorrect addition of scores
Loe, Kadlubek, & Marks (2007)	17 graduate students	WISC-IV	An average of 25.8 errors per protocol; common errors were failure to query, assigning too many points to a response, failure to record an examinee's response, and inaccurate test composite scores, resulting in incorrect FSIQ and Verbal Comprehension Index

Note. WISC = Wechsler Intelligence Scale for Children; WISC-R = Wechsler Intelligence Scale for Children–Revised; FSIQ = Full Scale IQ; WISC-III = Wechsler Intelligence Scale for Children, Third Edition; VIQ = Verbal IQ; WISC-IV = Wechsler Intelligence Scale for Children, Fourth Edition.

In the above scenarios, an upward score bias was illustrated. However, the intelligence testing literature suggests that errors can also be biased in a downward direction. According to the American Psychological Association Dictionary of Psychology, a **halo effect** is "*the tendency for a general evaluation of a person, or an evaluation of a person on a specific dimension, to be used as a basis for judgment of the person on another specific dimensions*." According to Sattler (2001), halo effects are known to occur during intellectual testing, particularly on open-ended response items. If an examiner has a perception of an individual as being very capable, the examiner may score responses in an upward (higher) biased direction. Conversely, examiners may display a downward (lower) scoring bias for individuals perceived as less capable (Sattler, 2001). The potential for positive or negative halo effects is most likely increased in high-stakes assessments. For example, Schlesinger (2003) describes the possibility of a downward scorng bias via the mechanism of "*malingering by proxy*."

Psychologists engaged in intelligence testing (and those who receive psychological reports) need to recognize that the underlying psychometric scaling of individual tests in certain intelligence batteries have not evolved from early classical test theory (CTT) methods to methods based on modern item response theory (IRT) which can allow (if the measurement technology is appropriately harnessed) for more precise scaling of the individual test scores summed to obtain

the global IQ. Intelligence batteries such as the DAS-II and WJ III provide IRT-based individual subtest scales that provide greater score specificity and sensitivity (i.e., they have more closely spaced tick marks on the underlying measurement rulers for each individual test). Although some intelligence tests may report the use of IRT (e.g., Rasch) item scaling during item development, selection and item ordering, a reading of their respective technical manuals often reveals that the potential increase in individual test scale precision (due to more dense tick marked mental measurement scales) is often not harnessed. Despite using IRT methods for item calibration, the individual subtests on some IQ batteries return to some variant of the *normalized ss 1-19 scale* (e.g., KABC-II; SB5). In one case, the SB5 appears slightly schizophrenic regarding the harnessing of the powers of IRT-based scale development. IRT (Rasch) methods were used extensively during SB5 item development and calibration and were used to provide an IRT-based *CSS* (change sensitive scale) to more precisely measure growth over time. However, the IRT-based scale precision was apparently discarded in favor of the more historical/traditional (and less precise) *ss 1-19* subtest scoring system for individual SB5 tests.

The issues raised here should give pause to psychologists who interpret intelligence tests. Psychologists must be extra diligent in the administration and scoring of <u>all</u> tests in <u>all</u> intelligence batteries, but may need to be more so with IQ batteries that provide individual test scores based on the older (historical and traditional) and less sensitive raw score to *ss 1-19 scale*. Minor scoring errors, if of a sufficient number and across enough subtests, in the context of the *ss 1-19 scale* (M=10; SD=3), can result in changes of the final *sum of scaled scores* large enough to produce changes in the final composite IQ *SS* reported. **And, these changes may be of sufficient magnitude to have real world consequences.** Given the potential real-world consequences (e.g., Atkins MR/ID death penalty decisions; eligibility for special education or SSA benefits) of IQ *SS* changes that may occur more frequently on IQ test batteries that use the *less sensitive ss 1-19 scale* (which, ironically, is a scale that is *more sensitive* to the effects of single raw score point changes that may occur due to test administration and scoring errors), these *ss-to-SS* conversion procedures should not be undertaken without full knowledge of the potential measurement issues. Psychologists need to "know thy instruments."

Given the above, I offer the following suggested guidelines:

1. If there is a professionally appropriate reason for a psychologist to convert individual subtest scores based on the *ss 1 to 19 scale* to an IQ *SS* scale, the psychologist should not report the specific point value associated with the exact standard score (e.g., *ss* = 6 converts to an IQ *SS* of 80), but should report to five point IQ *SS* range associated with the specific scale score (78 to 82 in the above example).
2. When interpreting IQ tests that use the *ss 1-19 scale* for individual tests, psychologists should stick with the original *ss* score values unless there is a good reason for converting the *ss* scores to the IQ *SS* scale.
3. When conducting research on individual subtests based on the *ss 1-19 scale*, the analysis should be based on the original standard score *1 to 19 scale*.

4. When conducting research that requires the examination of scores from individual tests that are based on the two different scales (*ss* and *SS*) discussed here, the only option available is to use the specific *SS* point value associated with each scaled score, and then address the possible impact of the imprecision in the score transformation in the results and discussion sections of the analysis. It is possible that there may be more elegant statistical solutions to address the unknown imprecision introduced by the *ss-to-SS* conversion process, but I leave it to my colleagues who possess greater statistical and methodological skills to articulate such procedures.
5. Psychological examiners should routinely review the administration and scoring directions for **all** intelligence tests they administer. As reflected in the scoring error literature cited above, it is not uncommon for both novice and experienced examiners to make enough administration and/or scoring errors that result in significant (and often large) changes in the final composite IQ *SS* score. While this problem is inherent in the administration and scoring of <u>all</u> intelligence test batteries, as demonstrated above, **the sensitivity of multiple raw score scoring errors producing noticeable changes in the final composite IQ *SS* score is greater with intelligence batteries that rely on the original Wechsler –like *ss 1-19 scale* metric.**[ii]
6. When IQ test scores are to be used in the context of strict guideline driven cut-scores, psychological examiners would be wise to double check all scoring and seriously consider having a knowledgeable colleague (who is also experienced with the same test) independently rescore the entire test record and correct any flagged errors and reconcile any disagreements in scoring decisions.

[<u>Conflict of interest statement</u> -- I, Kevin McGrew, am a co-author of the *Woodcock Johnson Battery-Third edition, WJ III*, which is a direct commercial competitor with the all intelligence batteries mentioned in this example. Also, the material included in this draft report do not necessarily reflect the opinions of the other WJ III coauthors or the publisher of the WJ III]

---

[i] I deliberately used the WISC-III scoring tables for the examples so not to divulge any potential information regarding the *sums of scaled score to IQ relationships* for batteries that are in current use (WISC-IV; WAIS-IV). I further masked this information by using X, Y and Z in place of the actual *sums of scaled scores* associated with each specific IQ SS from the WISC-III scoring tables.

[ii] This statement is based on more than the information presented in this brief report. The degree of precision or imprecision in the underlying scale of an individual test from an intelligence battery is due to a combination of test development procedures that are beyond the scope of this current brief report. Issues involved include, but are not limited to: (a) using classical versus modern test theory methods for test scaling (and not just item development and selection), (b) converting raw scores to an IRT-based equal interval scale which is then used as the basis for constructing standard score norms, and (c) differences in the application of continuous norming procedures that provide norm tables that span multiple months (e.g., 3 to 6 months) versus those that provide norms for each month of age.