

APPLIED PSYCHOMETRICS 101:

#10: “Just say no” to averaging IQ subtest scores

The publishers and authors of intelligence test batteries provide norm-based composite scores based on two or more individual subtests. In practice, clinicians frequently form hypotheses based on combinations of tests for which norm-based composite scores are not available. In addition, with the emergence of Cattell-Horn-Carroll (CHC) theory as the consensus psychometric theory of intelligence, clinicians are now more frequently “crossing batteries” to form composites intended to represent broad or narrow CHC abilities. Beyond simple “eye-balling” of groups of subtests, clinicians at times compute the arithmetic average of subtest scaled or standard scores (pseudo-composites). This practice suffers from serious psychometric flaws and can lead to incorrect diagnoses and decisions. The problems with pseudo-composite scores are explained and recommendations made for the proper calculation of special composite scores.

Joel Schneider

Illinois State University

Kevin McGrew

Institute for Applied Psychometrics (IAP)

“Just say no” to averaging IQ subtest scores

Joel Schneider

Illinois State University

Kevin McGrew

Institute for Applied Psychometrics (IAP)

Psychologists typically use more than one type of test to measure a specific cognitive ability (e.g., working memory capacity, general knowledge, or visualization). Relying on a single test is dangerous because it is hard to determine whether a person’s performance is due to the ability that the test is theoretically intended to measure or if the performance was influenced by some highly specific aspect of the test, such as the test format. Measuring the targeted ability with multiple tests of differing formats allows clinicians to distinguish between performance that is consistent across tests measuring the same ability and performance that differs significantly depending on the type of test format.

A *composite score* is formed by the sum (or weighted sum) of two or more tests, typically called *subtests*.¹ This sum is then transformed into a readily interpretable standard score (e.g., scaled scores, index scores, and T-scores). Not only is a composite score generally more reliable than the subtest scores it comprises, it is typically more strongly correlated with the theoretical construct it operationalizes than any of the individual subtests in the composite. In other words, composite scores are usually more reliable and more valid than the individual subtests used to calculate the composite.

Most of the composite test scores used by psychologists are suggested by the test’s publisher. For example, the Wechsler batteries include composite scores such as the Full Scale IQ or the Verbal Comprehension Index. In contemporary intelligence test batteries (e.g., WISC-IV, WAIS-IV, SB5, DAS-II, KABC-II, and WJ-III) composite interpretative indices are most often based on evidence from factor analytic research. That is, subtests which correlate more highly with each other than with other subtests in the test battery are assumed to measure a common construct. This construct is operationalized by a composite score formed from the highly correlated subtests.

Test publishers do not provide means to calculate composite scores for all possible combinations of subtests. Most combinations of subtests are not particularly useful but some combinations of subtest scores omitted by test publishers are of great interest to clinicians. For example, The WAIS-IV Working Memory Index is formed from Digit Span and Arithmetic. Many clinicians are reluctant to use the Arithmetic subtest as a measure of working memory capacity since it also measures quantitative reasoning (Phelps, McGrew, Knopik, & Ford, 2005). Replacing the Arithmetic subtest with the Letter-Number Sequencing subtest provides a composite score with less conceptual ambiguity. That is, a composite of Digit Span and Letter-Number Sequencing is a more direct measure of working memory capacity. Unfortunately, this composite is not among the composite scores included in the official WAIS-IV

¹ Most test batteries refer to individual tests as *subtests*, although this is not uniform. The WJ III battery refers to individual tests as *tests*, and includes some tests that are comprised of two to three *subtests*, which are smaller measures for which derived scores are not provided.

scoring procedures. Clinicians must go to extra efforts to calculate this new WAIS-IV working memory composite properly.

Informal composites (composite scores without official sanction from test publishers) have an important place in the development of clinical procedures for test interpretation. A major feature of Alan Kaufman's (1979) "intelligent testing" philosophy is the process of clinicians generating hypotheses about an individual's pattern of strengths and weaknesses, often based on combinations of two or more tests for which the published cognitive test battery does not provide norm-based scores. Although the clinical intelligence test interpretation literature sometimes provides statistical equations for generating an approximate standard score for a clinical grouping of individual tests, this typically has been the exception.

Although the practice is widely recognized as suboptimal, clinicians often create composites by simply averaging the subtests included in the informal composite. Sometimes the "averaging" is done with the just the eyes (e.g., "These three scores are in the low average range so the ability they all measure in common is also in the low average range."). Such practices may suffice for the early stages of the interpretation process when the clinician is simultaneously mentally juggling alternative explanations. At this formative stage in interpretation it would be burdensome to laboriously calculate each composite score considered. However, when the final interpretation is presented, the scores should be as accurate as possible. Unfortunately, many clinicians mistakenly believe that their averaging procedures are sufficiently accurate to spare them the effort of calculating true composite scores. Clinicians are often unaware of the counterintuitive features of composite scores, such as the fact that a composite score can be higher (or lower) than any of the subtests that comprise it, sometimes by a wide enough margin to alter the interpretation substantially.

Problems with Pseudo-Composite Standard Scores

One of the benefits of the adoption of the *CHC theory of intelligence* as the consensus psychometric framework by which contemporary cognitive ability test batteries are designed and interpreted (Keith & Reynolds, 2010; McGrew, 2009; Netwon & McGrew, 2010; Schneider & McGrew, 2011) is that clinicians are now able to combine information within and between test batteries to form theoretically informed composite scores. For example, the WISC-IV has only one subtest (Block Design) that is a strong and unambiguous measure of the CHC construct Gv (General Visual-Spatial Processing). Clinicians using the WISC-IV are able to select other strong and unambiguous measures of Gv in other batteries (e.g., KABC-II Triangles, DAS-II Pattern Construction, SB5 Nonverbal Visual-Spatial Processing, or WJ III Spatial Relations) to form a more reliable and valid composite score measuring Gv. Similarly, even within the same battery, users often find it necessary to create informal cluster composites. An example would be combining the WJ III NU Retrieval Fluency and Rapid Picture Naming tests into a composite score to estimate the speed at which semantic information stored in long-term memory can be accessed. Another reason to use informal composites within the same battery is to take advantage of relevant information from routinely administered supplemental subtests. For example, the WISC-IV Verbal Comprehension Index consists of the Similarities, Vocabulary, and Comprehension subtests. However, the WISC-IV includes two additional subtests, Information and Word Reasoning, that also measure crystallized intelligence (Gc). There is no good reason to use a three-subtest composite when a more reliable and valid five-subtest composite is available.

Although combining two or more individual tests within or across batteries is a defensible practice within the context of the clinical "intelligent" hypothesis formation, creating composites by averaging scores introduces unnecessary measurement error. For most individuals tested the measurement error is small

and does not result in serious distortion of the interpretation process. Unfortunately, most people referred for evaluations are not like the majority of the people in the population. They are typically referred for testing due to discrepant cognitive or achievement abilities. *Worse, the procedure of averaging subtest scores creates the most measurement error in the cases in which accuracy matters most: when the scores are extreme and there is a high-stakes decision to be made based on strict "bright line" eligibility or diagnostic classification criteria.*

The core problem with averaged *pseudo-composites* is that the average of individual subtest SS's *does not have the same standard deviation as the scores being averaged*. The standard deviation is different because averaged pseudo-composites fail to account for the intercorrelations of the subtests being averaged. Scores with different standard deviations are not on the same scale and are not directly comparable. The problem is particularly insidious because the two types of scores seem to fit together naturally and thus, the interpretive errors caused by averaged pseudo-composites are easily overlooked.

The mathematics of composite scores was worked out over a century ago by a number of statisticians, including Spearman (1904). The problem inherent in averaged pseudo-composite scores was vividly illustrated in an influential article in the *Journal of Consulting Psychology* by Tellegen and Briggs (1967). The lasting influence of this article may have something to do with its memorable title "*Old wine in new skins: Grouping Wechsler subtests into new scales.*" Tellegen and Briggs presented formulas for calculating psychometrically correct informal composite scores (see the *Appendix* for a more user-friendly formula). These equations properly account for the subtest correlations (and thus, indirectly the reliability of subtest scores) so that the composite score is converted into a readily interpretable standard score.

Problems with Averaged Pseudo-Composites: Real-Word Demonstration

Rather than first presenting mathematical explanations and equations that demonstrate the problems with averaged pseudo-composites, the problem is demonstrated here with plots of "real" norm-based composites and averaged pseudo-composites derived from the WJ III NU standardization sample (McGrew & Woodcock, 2001; Woodcock, McGrew, Schrank, & Mather, 2001, 2007).

The following steps were implemented for each specific composite score comparison. These steps were completed with the subjects from the WJ III NU standardization sample. Although these steps refer to the Gv composite, analogous procedures were implemented with other composites.

1. All members of the WJ III NU standardization sample were identified for whom age-based Visual Processing (Gv) cluster scores were available.
2. A pseudo-composite for the Gv was calculated from the arithmetic average of each norm subject's Spatial Relations and Picture Recognition SS's. That is,

$$\text{Pseudo-Gv} = \frac{\text{Spatial Relations SS} + \text{Picture Recognition SS}}{2}$$

3. All WJ III NU norm subjects were sorted from lowest to highest based on their real norm-based Gv SS. Subjects were then grouped into successive blocks of 50 subjects each. For each block of 50 subjects, the mean real Gv scores and pseudo-Gv scores were calculated. This reduction of the number of data points was completed to make the graphs easier to interpret.

4. Each subject's pseudo-Gv score was subtracted from the subjects' respective real Gv score. These real-pseudo-Gv difference scores were plotted as a histogram (See Figure 1).
5. The real and pseudo-Gv scores (for each block of 50 sorted subjects) values were plotted with the real Gv values on the x-axis and the pseudo-Gv values on the y-axis. The distance-weighted smoother was then applied to the data points. A linear (1-1) function line was then superimposed on the plot.

Because both the real and pseudo-Gv scores are approximately normally distributed, the difference between them is also approximately normally distributed (See Figure 1). The mean difference score is approximately zero (0.3) and the SD of the difference score distribution is 3.2 points. This means that about 68% of the WJ III NU norm sample subjects displayed differences between their real/pseudo-Gv SS's from approximately -3.2 to +3.2 (*a range of 6.4 SS difference points*). About 95 % of the WJ III NU norm sample subjects displayed real/pseudo-Gv SS differences from approximately -6.4 to +6.4 (*a range of approximately 13 SS difference points*). The real/pseudo-Gv difference scores ranged from approximately -14 points (-14.2) to +15 points—*a range of 29 points!*

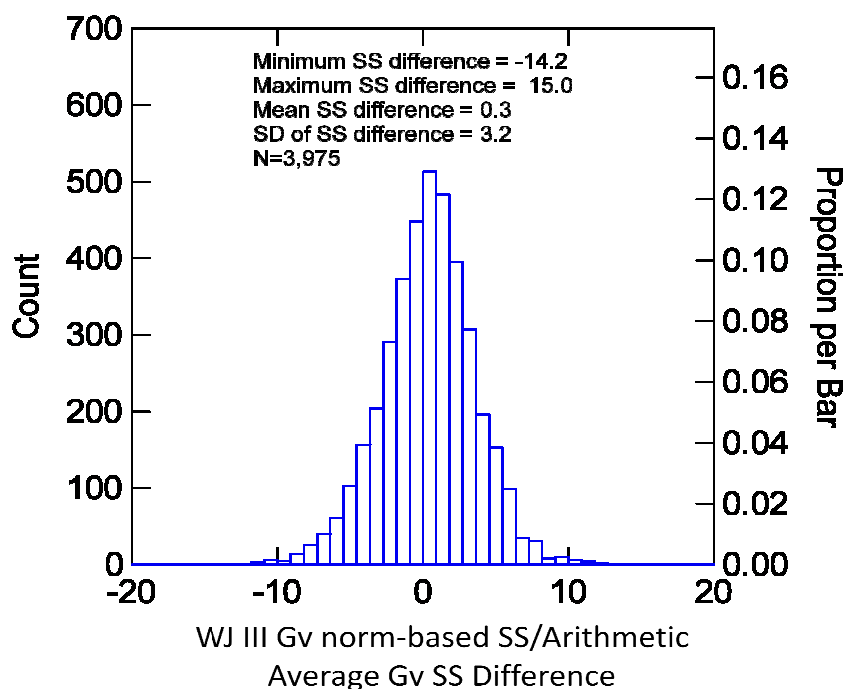


Figure 1: Histogram of WJ III the SS Differences between Real and Pseudo-Gv Composites

The fact that the mean difference between the real and pseudo-Gv scores is approximately zero may, at first glance, suggest that pseudo-Gv is an unbiased estimate of real Gv. Not so! *The differences between the real and pseudo-Gv SS differences vary systematically as a function of ability level on the construct of interest (Gv).* Figure 2 presents the information described in # 5 above. If the simple arithmetic average of Spatial Relations and Picture Recognition SS's were an accurate score, as represented by their real norm-based Gv scores, one would expect the plotted values to fall on the perfect linear regression line

superimposed on the plot. Instead, one observes a near linear smoothed function that rotates at the real Gv SS (x-axis) and pseudo-Gv SS (y-axis) pivot point (where both equal 100). Values to the left-side of the pivot point are above the desirable linear target line. This means that the *differences between real and pseudo-Gv SS's systematically (become higher than the real score) the lower a subject scores on the real Gv score distribution*. Conversely, the opposite trend is observed to the right of the pivot point. *The higher a person's real Gv SS the greater the pseudo-Gv SS diverges from the real score, in the direction of being an underestimate.*

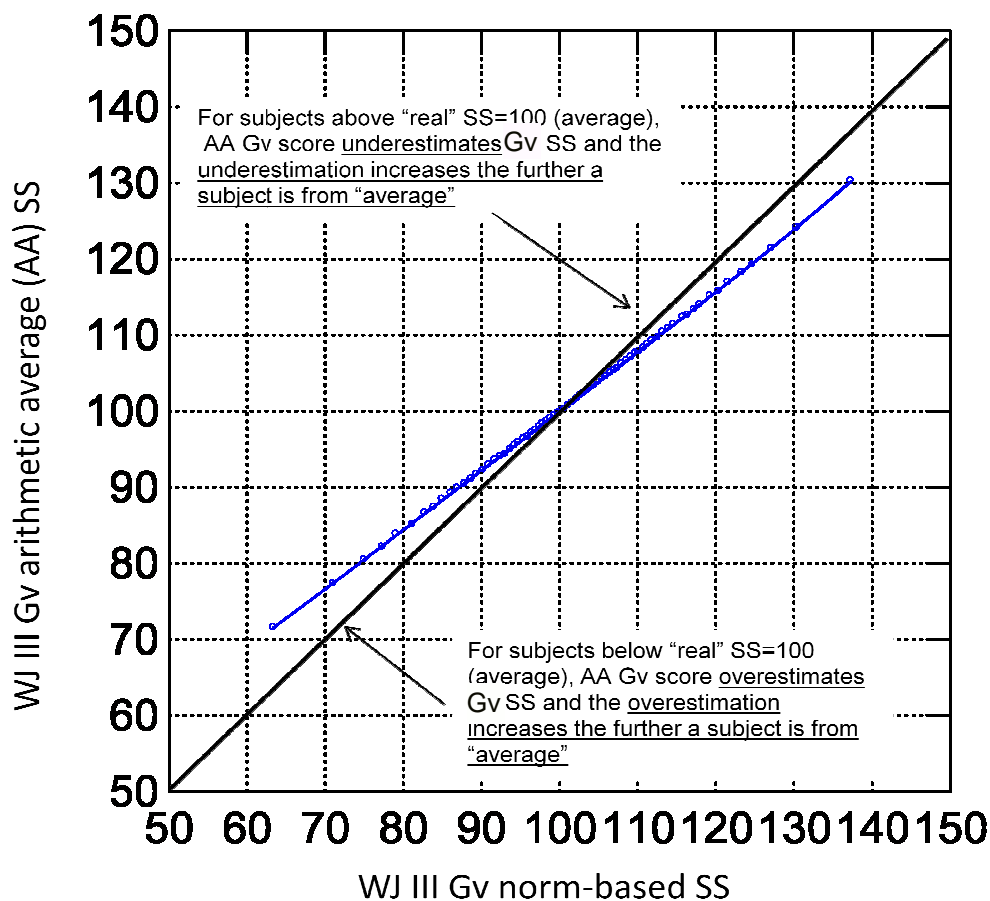


Figure 2: Plot of Real and Pseudo-Gv Composite Scores

The results presented in Figures 1 and 2 leads to the obvious conclusion that *averaged pseudo-composite scores are not accurate estimates of an individual's real norm-based composite scores*.

Why?

The fact that real composites are more extreme (i.e. further from the population mean) than averaged pseudo composites is difficult to grasp. Broadly speaking, it occurs because it is more unusual to be extreme on multiple dimensions than it is to be extreme on only one dimension. For example, there are few extremely tall people. There are few people with extremely good hand-eye coordination. There are

even fewer people who are extremely tall AND have extremely good hand-eye coordination. If the two variables were combined into a composite score (perhaps as a predictor of basketball talent), a person who is three standard deviations above the mean on both variables would be higher than three standard deviations of the mean on the composite score because the combination of two rare traits is very rare indeed. Consider what it would mean in this case if the composite score equaled the averaged pseudo-composite score. It would mean that all of the people in the top percentile of height are also in the top percentile of hand-eye coordination, something that is plainly not in accordance with reality.

There are three factors that influence the discrepancy between real and pseudo-composites. The **first**, mentioned previously, is that the more extreme an individual's subtest scores are (i.e., the more distant from the population mean), the more inaccurate the averaged pseudo-composite score becomes.

The **second** factor that influences the inaccuracy of averaged pseudo-composites was noted by McGrew and Flanagan (1998):

The failure of a normed cluster score to equal the arithmetic average of the individual tests that make up the cluster is related directly to the magnitude of the intercorrelations and the number of tests in the cluster (Paik & Nebenzahl, 1987). The lower the intercorrelations between the tests that contribute to a cluster score, the more extreme will be the difference between a cluster score that is 'normed' and a cluster that is based on the arithmetic average of tests. The only time that a cluster standard score will equal the average of the individual tests that make up the cluster is when all the subtests of the cluster are correlated perfectly (see Paik & Nebenzahl, 1987, for details) (p. 413).

Figure 3, which is the comparable figures for the WJ III NU Gf (Fluid Reasoning) composite score, demonstrates the importance of incorporating information regarding the known correlation between measures in the appropriate calculation of a valid averaged pseudo-composite score.

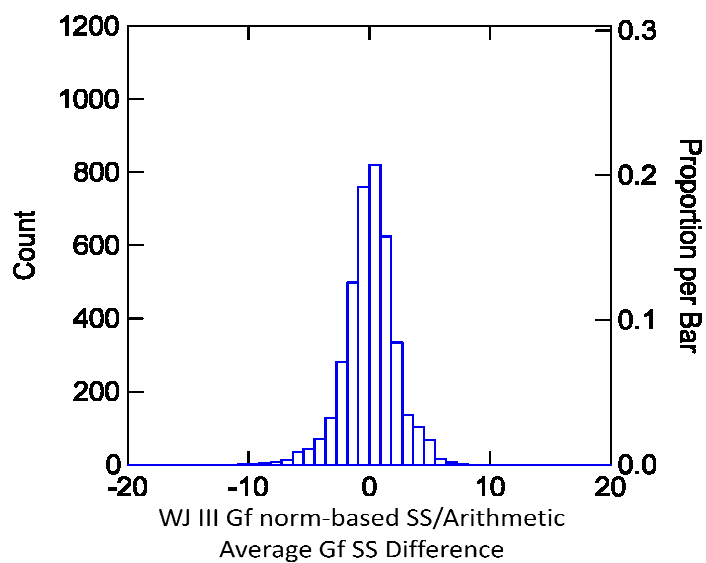


Figure 3: Histogram of WJ III the SS Differences between Real and Pseudo-Gf Composites

The range of the differences between real and pseudo-composite scores (Figure 3) and the amount of over- and under-estimation (as a function of a person's real Gf score, see Figure 4) is considerably *smaller* than for the Gv composite (see Figure 4). Why?

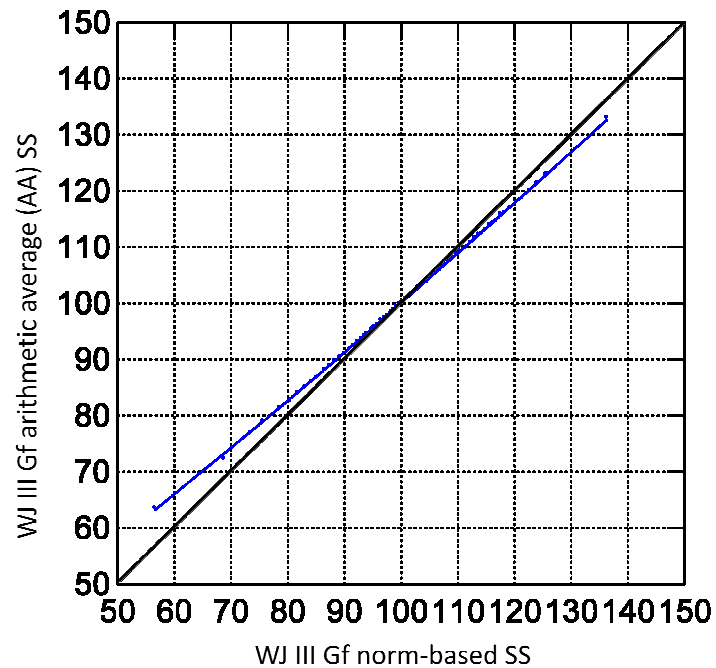
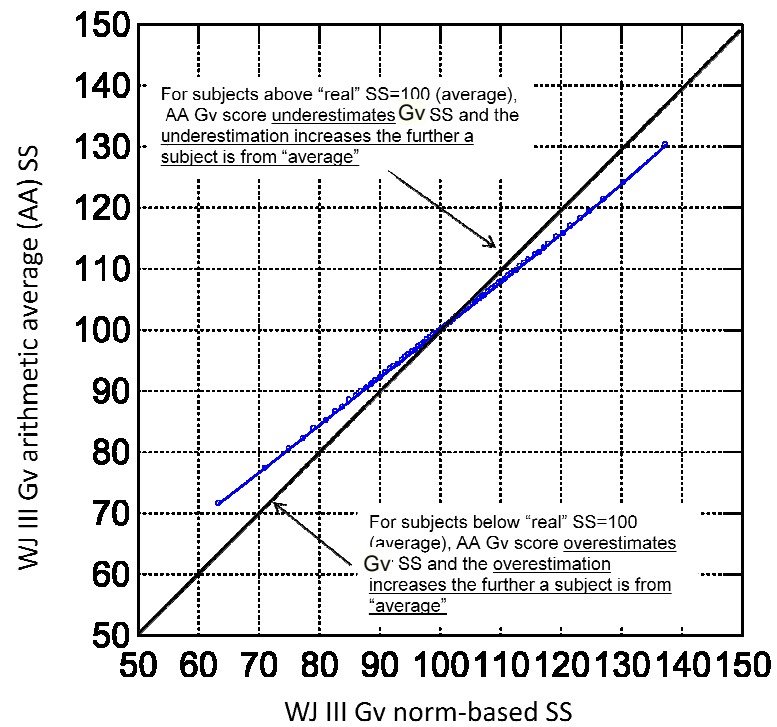


Figure 4: Comparison Plots of Real and Pseudo-Gf (bottom graph) and Pseudo-Gv (top graph) Composite Scores

The correlation across the complete WJ III NU norm sample between the two tests included in the WJ III NU Gf cluster (Concept Formation and Analysis-Synthesis) is .56, while the correlation between the two tests included in the WJ III NU Gv cluster is .17. *The weaker the correlation between tests that comprise a composite the greater will be the range of real/pseudo-composite differences and the greater the degree of over- (below average) and under-estimation (above average) that will be present in the averaged pseudo-composite.*

The **third** factor that influences the inaccuracy of averaged pseudo-composites is the number of subtests in the composite. *The more subtests in the composite, the more inaccurate the averaged pseudo-composite.*² For example, as seen in Figures 5 and 6, the 7-subtest averaged pseudo-composite analogous to the WJ III NU General Intellectual Ability (GIA) cluster, is even more inaccurate than the 2-subtest averaged pseudo-composites in Figure 4.

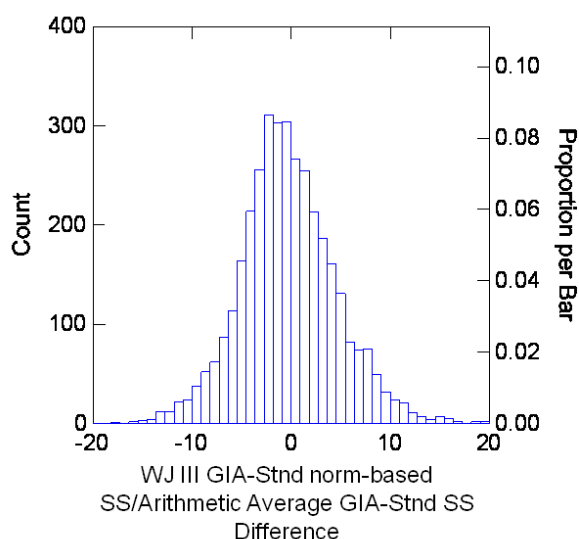


Figure 5: Histogram of WJ III the SS Differences between Real and Pseudo-GIA-Stnd Composites

² This is a very bad reason to avoid composites with many subtests! They are typically more reliable and valid than composites with fewer subtests.

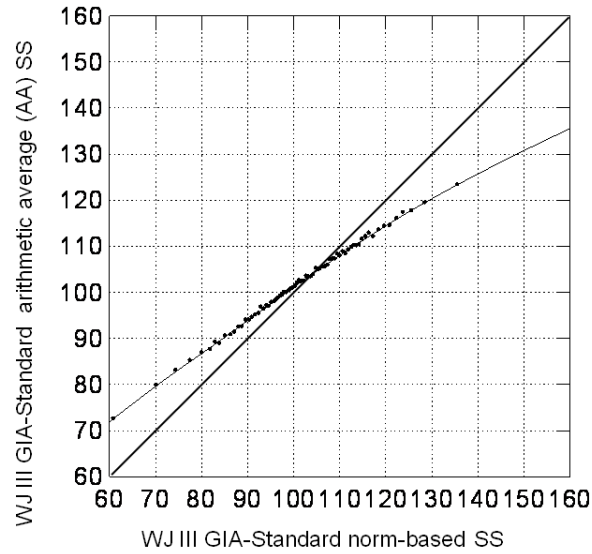


Figure 6: Plot of Real and Pseudo-GIA Composite Scores

In summary, if C is an individual's composite score and \bar{S} is the corresponding averaged pseudo-composite score, the difference between them ($C - \bar{S}$) is influenced by three factors.

1. $(\bar{S} - \mu)$ = The extremity of the pseudo-composite score (i.e., its distance from the population mean μ)
2. \bar{r} = The average intercorrelation between the subtests in the composite.
3. k = The number of subtests in the composite

As shown in the *Appendix*, these three factors influence the inaccuracy of an averaged pseudo-composite, as can be seen in this formula:

$$C - \bar{S} = (\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + \bar{r}(k-1)}} - 1 \right)$$

Figure 7 shows which combinations of k and r will make an averaged pseudo-composite inaccurate by 5 index score points or more.

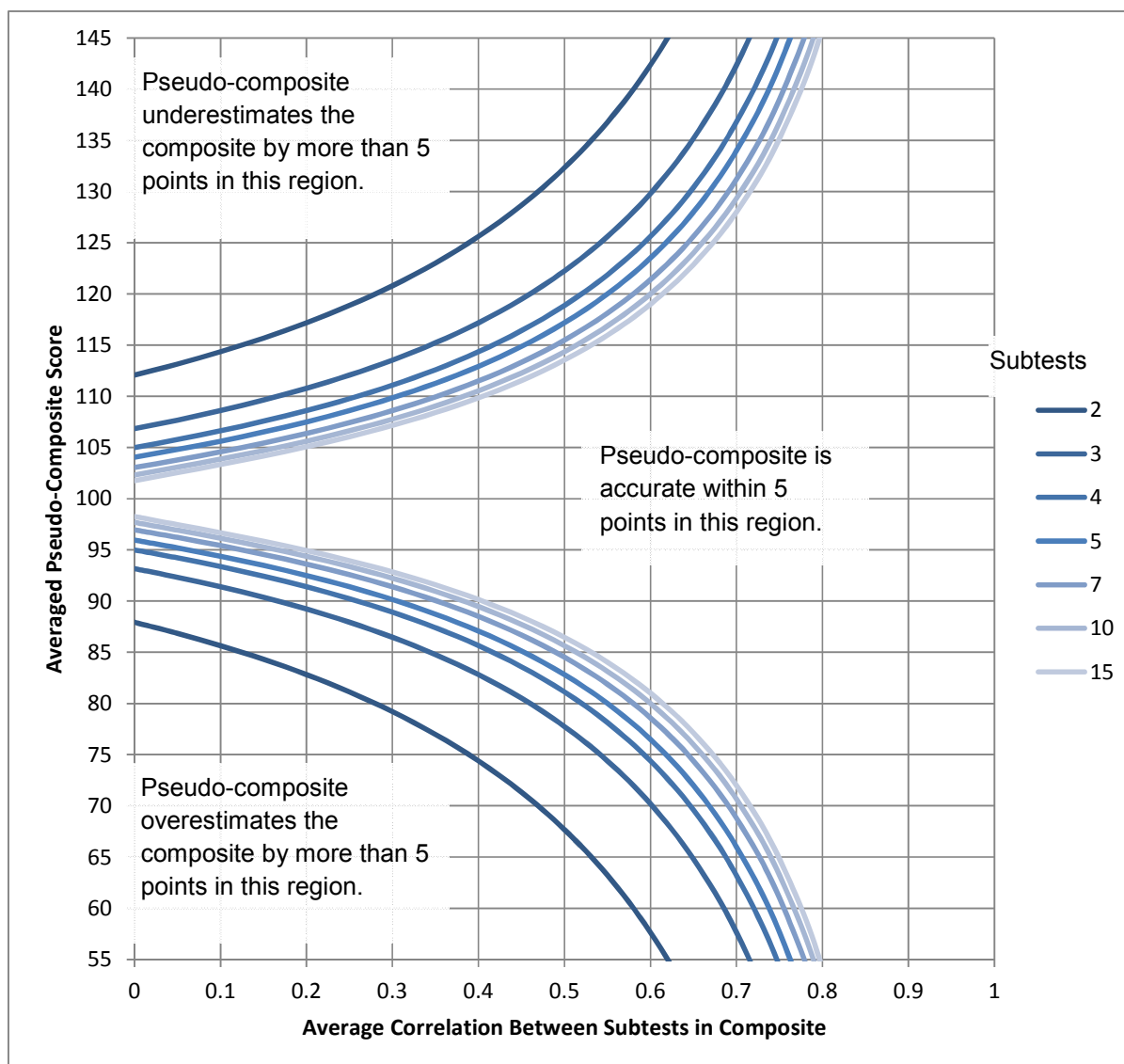


Figure 7: Combinations of Average Subtest Intercorrelations and Number of Subtests in the Composite that Produce Inaccurate Averaged Pseudo-Composite Scores

Figure 8 shows the factor by which the averaged pseudo-composite score is discrepant from the real composite score. For example, the figure shows that for an 8-subtest composite score in which the subtests have an average correlation of 0.3, the composite score is 1.6 times more extreme than the averaged pseudo-composite score (in terms of both scores deviations from the population mean). Suppose that the averaged pseudo-composite is 80. Its deviation from the population mean is $80 - 100 = -20$. The composite score's deviation is 1.6 times -20, which equals -32. Adding the population average (mean) back again yields a composite score of 68. Thus, the averaged pseudo-composite score is inaccurate by $80 - 68 = 12$ points.

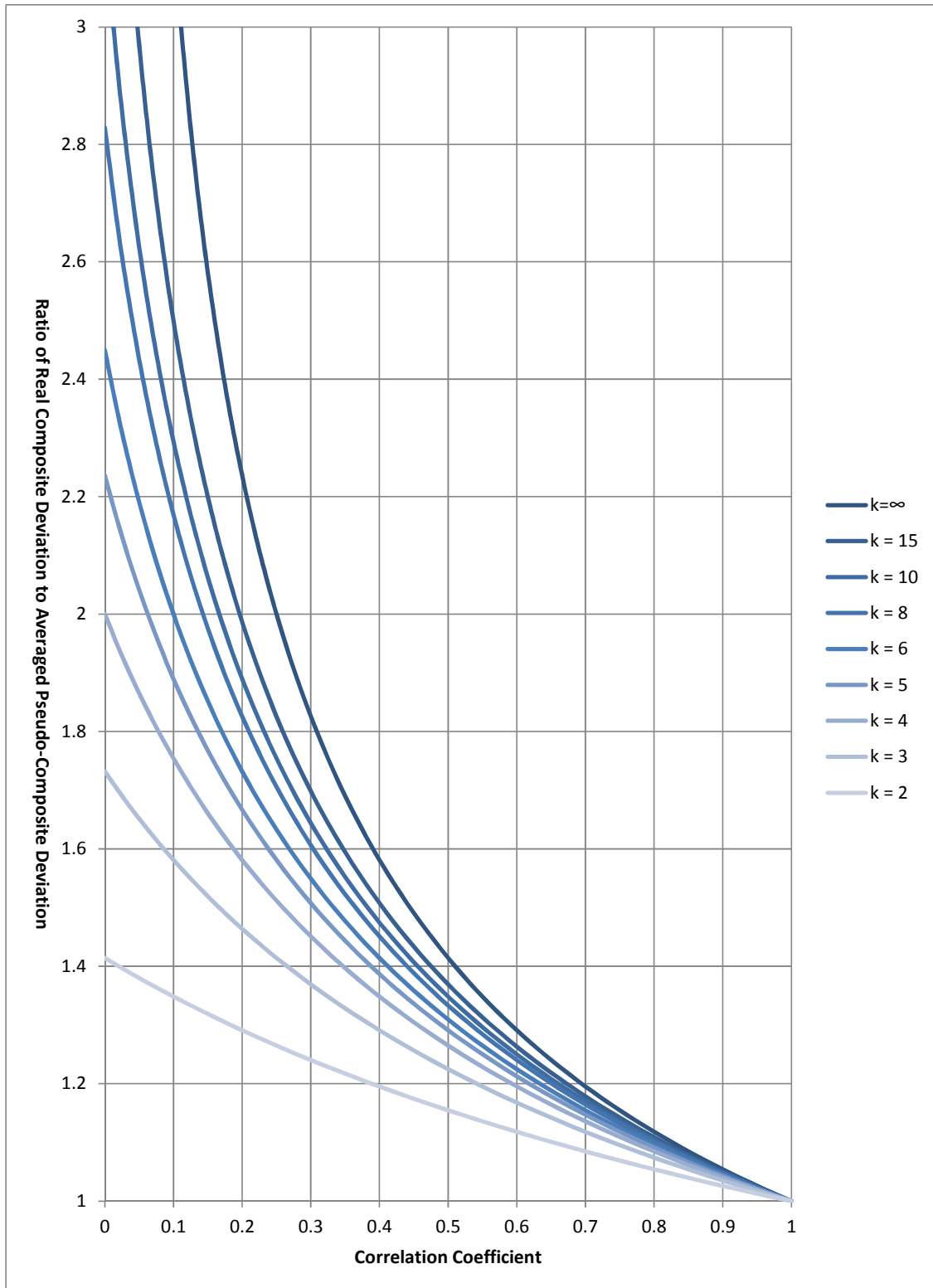


Figure 8: The Factor by Which Averaged Pseudo-Composite Deviations Differ from Real Composite Deviations as a Function of the Average Intersubtest Correlation and the Number of Subtests in the Composite Score

A Complimentary Explanation of “Why?”

It is sometimes helpful to understand a phenomenon in two different modalities, especially when the phenomenon resists accurate conceptualization (as is the case here). It is not intuitive that a composite score will be more extreme than its accompanying averaged pseudo-composite score. It is also not immediately apparent that the degree of difference between the two scores decreases with subtests that are more correlated. To help guide the intuition to grasp the nature of the phenomenon more accurately, a geometric representation of composite scores is presented here.

Imagine that Subtest X and Subtest Y form a composite score. Subtest X 's score is plotted on the X -axis and Subtest Y 's score is plotted on the Y -axis. In Figure 9, Subtest X and Subtest Y both equal 130. The lengths of the red and blue vectors represent the deviations from the population mean of 100. If subtests X and Y are uncorrelated, the X - and Y -axes are orthogonal, meaning that they meet at a right angle. However, if subtests X and Y are correlated, we must rotate the Y -axis so that the angle (θ) between the axes is equal to the arccosine of the correlation. That is, $\cos(\theta) = r_{XY}$. This may seem like an odd thing to do but it is often done by statisticians dealing with correlated variables (e.g., in factor analysis, oblique rotations are quite common).

The length of the purple vector is equal to how far the real composite score is from the population mean of 100. Because subtests X and Y are equal in this case, their scores also equal the averaged pseudo-composite. Thus, it can be seen that as the correlation between X and Y increases from 0 to 0.9, the length of the purple vector becomes more like the deviation of the averaged pseudo-composite from the population mean of 100. Specifically, when the subtest scores are 130, the composite score is 142 when the correlation is 0 and 131 when the correlation is 0.9.

Essentially what is happening is that the sum of X and Y is being rescaled to the same metric as the X - and Y -axes. *This is precisely what a composite score is: a transformed sum that is interpreted on the same scale as its parts.*

There are a number of other very interesting phenomena that are elucidated by a geometric approach to composite scores—where difference scores are located in the graph, the analogy to oblique rotations in factor analysis, and the generalization of the two-dimensional plots depicted here to multidimensional space. However, we will not explore these issues here.

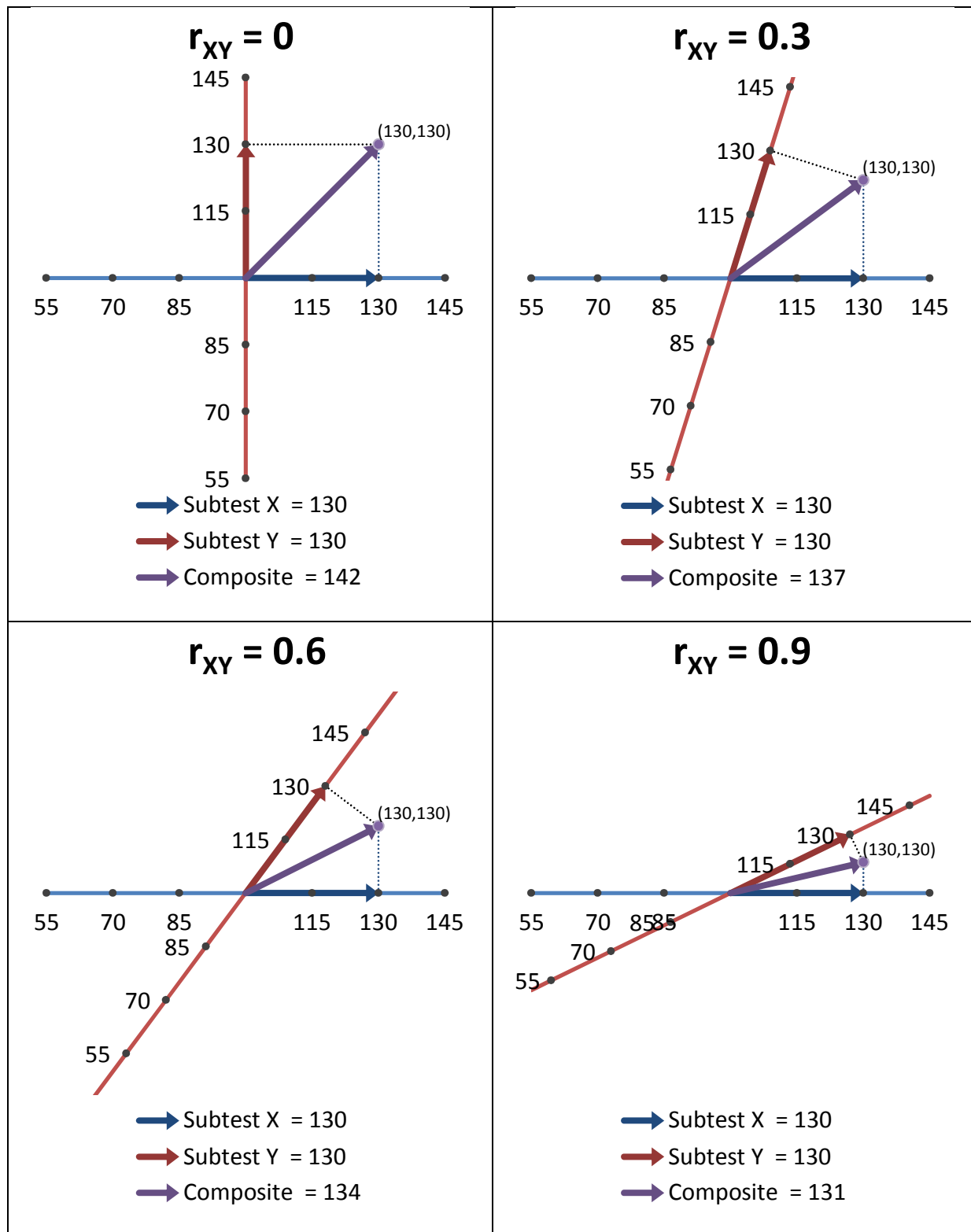


Figure 9: A Geometric Representation of How the Correlation between Subtests Alters the Relationship between Subtest Scores and Composite Scores

Comments on the “Only-When-Less-Than-1-SD” rule-of-thumb solution

Finally, some have suggested an informal rule to limit the potential error in the calculation of averaged pseudo-composites: only create pseudo-composites when the tests to be averaged are relatively consistent. The rule-of-thumb heard from the field is that it is appropriate to average test SS's if the SS's for the respective tests are within 15 SS (< 1 SD) points of each other.

The proof that this rule-of-thumb is not appropriate or supported is evident when one inspects the formula of the difference between the real composite and the pseudo-composite ([see bottom of page 10](#)). The magnitude of the difference depends on (a) the average of the subtests, (b) the number of subtests, and (c) the average correlation between the subtests. *Differences or discrepancies among the subtests are nowhere to be seen in the formula and thus, are irrelevant to the issue at hand.*

Even though the rule-of-thumb provides no protection from the measurement error introduced by averaged pseudo-composites, it might seem that the rule is still useful if it prevents clinicians from misinterpreting test data. It has been argued that, when there is within-composite scatter in the subtest scores, the composite is “not interpretable” or less valid, which presumably means that it is less correlated with the theoretical construct the composite is intended to measure. Although this idea has intuitive appeal, it is not true. Although the mathematical basis of this conclusion is beyond the scope of this paper, composite scores are just as valid when subtests are discrepant as when they are consistent. It will suffice to say that construct irrelevant influences are just as likely to produce consistent subtest scores (when their influence moves the subtest scores in the same direction) as they are to produce discrepant subtest scores (when their influence moves the subtest scores in opposite directions). Thus, when subtests are consistent, there is no increased assurance that the composite score is more accurate or more valid. Likewise, a large subtest discrepancy, in and of itself, does not signal that the composite score is less accurate or less valid. Indeed, the whole point of making composite scores is the hope that construct-irrelevant influences cancel each other out.

Recommendations and Solutions

- Averaged pseudo-composite scores should **not** be used when making critical decisions about individuals (e.g., program eligibility; diagnosis or classification) particularly in contexts where specific scores are compared to “bright line” specific cut-score eligibility/diagnostic criteria. Only real norm-based scores possess the necessary psychometric accuracy for this purpose. *Under no circumstances should averaged pseudo-composite scores be entered into equations, formulas, or procedures that involve high-stakes and important decisions regarding individuals.*
- Fortunately, more-and-more contemporary IQ batteries are now explicitly or implicitly designed as per CHC theory, or have been found to best be interpreted from this psychometric framework (Keith & Reynolds, 2010). Thus, practitioners are now provided with more norm-based composite scores based on valid psychometric theory. It is strongly recommended that assessment professionals use the norm-based composite scores provided within each respective IQ battery. If supplementary testing is required (crossing batteries) to obtain at least a two-test composite of an ability, *it is recommended that a supplementary battery be selected that provides 2 or more test normed-based composite SS*. That is, instead of taking one test from a core battery and combining it with the score from a single test from a different cognitive battery, it is preferable to select two or more tests from a supplementary battery that will produce a norm-based composite ability score—mitigating the need to engage in psychometrically risky cross-battery composite score generation.

- *If informal composite scores are calculated, users should use the available equations (see the Appendix and the references in this manuscript) to generate psychometrically sound scores.* If a composite is derived from within a single IQ battery, the respective technical manual should include the necessary psychometric information (e.g., correlations between tests to comprise a composite) to insert into the proper formula's. If such information is not available (e.g., a composite is created from Test 1 from Battery A and Test 2 from Battery B), then users should attempt to locate published research that would provide the best estimate of the correlation between Test 1 and 2. The reliability information for each test should be available in each batteries respective technical manual. Given the ease with which individuals can craft spreadsheets (e.g., Excel) and spreadsheet templates that include the required equations, it is not unreasonable to expect those who engage in this practice to create or locate such a tool (e.g., the first author's [Composite Score Calculator](http://my.ilstu.edu/~wjschne/CompositeScores.xlsx) ³). Even better yet is for test publishers or other agencies to provide software (that incorporates the technical characteristics of tests in a battery—reliabilities; correlations) to allow for the calculation of psychometrically defensible informal composite scores within the battery. One example is the first author's [Compositator](http://www.woodcock-munoz-foundation.org/press/compositator.html) (Schneider, 2010 ⁴) that works with the WJ III NU battery.

³ <http://my.ilstu.edu/~wjschne/CompositeScores.xlsx>

⁴ <http://www.woodcock-munoz-foundation.org/press/compositator.html>

References

- Flanagan, D. P., & McGrew, K. S. (1998). Interpreting intelligence tests from contemporary *Gf-Gc* theory. Joint confirmatory factor analyses of the WJ-R and KAIT in a non-white sample. *Journal of School Psychology, 36*, 151-182.
- Kaufman, A. S. (1979). *Intelligent Testing with the WISC-R*. New York: Wiley & Sons.
- Keith, T. Z., & Reynolds, M. R. (2010). Cattell–Horn–Carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools, 47*(7), 635–650.
- McGrew, K. S. (2009). Editorial: CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence, 37*, 1–10.
- McGrew, K., & Flanagan, D. (1998). *The Intelligence Test Desk Reference (ITDR). Gf-Gc cross-battery assessment*. Boston. Allyn & Bacon.
- Newton, J. H., & McGrew, K. S. (2010). Introduction to the special issue: Current research in Cattell-Horn-Carroll-based assessment. *Psychology in the Schools, 47*(7), 621–634.
- Phelps, L., McGrew, K. S., Knopik, S. N., & Ford, L. (2005). The general (*g*), broad, and narrow CHC stratum characteristics of the WJ III and WISC-III tests. A confirmatory cross-battery investigation. *School Psychology Quarterly, 20*(1), 66-88.
- Schneider, W. J. (2010). *The Compositor 1.0*. WMF Press.
- Schneider, W. J., & McGrew, K. (in preparation) The Cattell-Horn-Carroll model of intelligence. To appear in D. Flanagan & P. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (3rd ed.). New York: Guilford.
- Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology, 15*, 201-293.
- Tellegen, A., & Briggs, P. (1967). Old wine in new skins: Grouping Wechsler subtests into new scales. *Journal of Consulting Psychology, 31*, 499-506.
- Woodcock, R., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson III*. Itasca, IL. Riverside.
- Woodcock, R., McGrew, K., Schrank, F., & Mather, N. (2001, 2007). *Woodcock-Johnson III Normative Update*. Rolling Meadows, IL. Riverside Publishing.

Appendix

Assume that there are k subtest scores, $S_1, S_2, S_3, \dots, S_k$. Each subtest has the same population mean μ and standard deviation σ . The average of all the intercorrelations between the subtest scores is \bar{r} . An averaged pseudo-composite score \bar{S} is formed by averaging all the subtest scores.

$$\bar{S} = \frac{\sum_{i=1}^k S_i}{k}$$

A real composite score C is formed when the sum of these subtests is transformed to have the same mean and standard deviation.

$$C = \frac{k(\bar{S} - \mu)}{\sqrt{k + \bar{r}k(k-1)}} + \mu$$

Note that $k + \bar{r}k(k-1)$ is equal to the sum of all elements (including the 1's in the diagonal) in the correction matrix of the subtests. That is,

$$k + \bar{r}k(k-1) = \sum_{i=1}^k \sum_{j=1}^k r_{ij}$$

Where r_{ij} is the correlation between subtest i and j and $r_{ij} = 1$ when $i = j$.

Also note that $k(\bar{S} - \mu)$ is equal to the sum of the deviations of each subtest from the population mean. That is,

$$k(\bar{S} - \mu) = \sum_{i=1}^k (S_i - \mu)$$

The difference between the real and the averaged pseudo-composite is

$$C - \bar{S} = \frac{k(\bar{S} - \mu)}{\sqrt{k + \bar{r}k(k-1)}} + \mu - \bar{S}$$

Simplifying and reorganizing yields

$$\begin{aligned} C - \bar{S} &= \frac{k(\bar{S} - \mu)}{\sqrt{k + \bar{r}k(k-1)}} - (\bar{S} - \mu) \\ C - \bar{S} &= (\bar{S} - \mu) \left(\frac{k}{\sqrt{k + \bar{r}k(k-1)}} - 1 \right) \\ C - \bar{S} &= (\bar{S} - \mu) \left(\frac{\sqrt{k}\sqrt{k}}{\sqrt{k}\sqrt{1 + \bar{r}(k-1)}} - 1 \right) \\ C - \bar{S} &= (\bar{S} - \mu) \left(\frac{\sqrt{k}}{\sqrt{1 + \bar{r}(k-1)}} - 1 \right) \\ C - \bar{S} &= (\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + \bar{r}(k-1)}} - 1 \right) \\ C - \bar{S} &= (\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + \bar{r}k - \bar{r}}} - 1 \right) \end{aligned}$$

When $\bar{r} = 0$,

$$C - \bar{S} = (\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + 0k - 0}} - 1 \right)$$

$$C - \bar{S} = (\bar{X} - \mu)(\sqrt{k} - 1)$$

When $\bar{r} = 1$,

$$C - \bar{S} = (\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + 1k - 1}} - 1 \right)$$

$$C - \bar{S} = (\bar{S} - \mu) \left(\sqrt{\frac{k}{k}} - 1 \right)$$

$$C - \bar{S} = (\bar{X} - \mu)(\sqrt{1} - 1)$$

$$C - \bar{S} = (\bar{S} - \mu)(1 - 1)$$

$$C - \bar{S} = (\bar{S} - \mu)0$$

$$C - \bar{S} = 0$$

When the number of subtests (k) increases, the difference between C and \bar{S} approaches

$$\lim_{k \rightarrow \infty} (C - \bar{S}) = \lim_{k \rightarrow \infty} \left((\bar{S} - \mu) \left(\sqrt{\frac{k}{1 + \bar{r}k - \bar{r}}} - 1 \right) \right)$$

$$\lim_{k \rightarrow \infty} (C - \bar{S}) = (\bar{S} - \mu) \left(\sqrt{\frac{1}{\bar{r}}} - 1 \right)$$

The geometric representation of the 2-subtest composite consisting of subtests X and Y suggests the following alternate formula for a real 2-subtest composite:

$$C = \left(\frac{X + Y}{2} - \mu \right) \csc \left(\frac{\cos^{-1} r_{XY}}{2} \right) + \mu$$

When there are k subtests, the composite score is the distance from the intersection of all k subtest scores in k -dimensional oblique space (such that the k axes intersect at angles equal to the cosines of the correlations of the respective subtests) to the subspace (having $k-1$ dimensions) that is perpendicular to the centroid of k unit vectors, each of which is parallel to one of the k axes and starting at the origin.

Second author information and conflict of interest disclosure

Dr. Kevin S. McGrew, Ph.D., is an Educational Psychologist with expertise and interests in applied psychometrics, intelligence theories and testing, human cognition, cognitive and non-cognitive individual difference variables impacting school learning, models of personal competence, conceptualization and measurement of adaptive behavior, measurement issues surrounding the assessment of individuals with disabilities, brain rhythm and mental timing research, and improving the use and understanding of psychological measurement and statistical information by professionals and the public. Prior to establishing IAP, Dr. McGrew was a practicing school psychologist for 14 years. McGrew received his Ph.D. in Educational Psychology (Special Education) from the University of Minnesota in 1989.

Dr. McGrew is currently Director of the *Institute for Applied Psychometrics* (IAP), a privately owned applied research organization established by McGrew. He is also the *Research Director for the Woodcock-Munoz Foundation* (WMF), Associate Director for *Measurement Learning Consultants* (MLC), and a *Visiting Professor in Educational Psychology* (School Psychology) at the University of Minnesota.

Dr. McGrew coauthored the current document in his role as the Director of IAP. The opinions and statements included in this report do not reflect or represent the opinions of WMF, MLC, or the University of Minnesota.

More complete professional information, including his professional resume, bio, and conflict of interest disclosures can be found at each of his three professional blogs and web page:

- www.iqscorner.com
- www.atkinsmrdeathpenalty.com
- www.ticktockbraintalk.blogspot.com
- www.iapsych.com