Woodcock Johnson® IV

# Assessment Service Bulletin Number 7

## Why Are WJ IV™ Cluster Scores More Extreme Than the Average of Their Parts? A Gentle Explanation of the Composite Score Extremity Effect

*W. Joel Schneider, PhD*

*Composite scores are more extreme than the average of the test scores used to compute them. Because few other kinds of measurement behave this way, it is a frequent source of confusion to assessment professionals. Several analogies and technical explanations are offered to help readers understand this phenomenon.*

Houghton
Mifflin
Harcourt

Manufactured in the United States

**Reference Citation**

- To cite this document, use:

  Schneider, W. J. (2016). *Why Are WJ IV Cluster Scores More Extreme Than the Average of Their Parts? A Gentle Explanation of the Composite Score Extremity Effect* (Woodcock-Johnson IV Assessment Service Bulletin No. 7). Itasca, IL: Houghton Mifflin Harcourt.

For technical information, please visit http://www.wj-iv.com or call HMH Customer Experience at 800.323.9540.

# Why Are WJ IV™ Cluster Scores More Extreme Than the Average of Their Parts? A Gentle Explanation of the Composite Score Extremity Effect

Our intuitions about measurement are shaped by everyday activities such as baking, home repairs, and exercise. The instruments we use—measuring spoons, rulers, and stopwatches—are not perfect, but they are quite reliable when used properly. Usually a single measurement settles the matter. The carpenter's adage—measure twice, cut once—is a warning not about fallible tools but about our own inconsistency and distractibility.

## Composite Scores Are Strange

Things are not so straightforward with *composite scores*, which are variables that summarize multiple test scores. Composite scores often behave in ways that violate our intuitions. For example, if a person scores 70 on both visual processing (*Gv*) tests of the *Woodcock-Johnson® IV Tests of Cognitive Abilities* (WJ IV COG; Schrank, McGrew, & Mather, 2014), Visualization and Picture Recognition, we would expect that the composite, the Visual Processing (*Gv*) cluster score, would also be 70. In truth, the *Gv* cluster score is around 65 (62 to 67, depending on the examinee's age and pattern of scores).[1] This counterintuitive feature of composite scores is also observed in the opposite direction: two high scores resulting in a cluster score that is even higher. For example, scores of 119 and 121 on two tests might result in a cluster score of 124.

This phenomenon, hereinafter referred to as the *composite score extremity effect*, refers to the counterintuitive fact that a composite score consisting of imperfectly correlated scores is always more extreme—further from the population mean—than the average of its parts. In Figure 1 on page 2, each point represents a person from the WJ IV standardization sample. Each person's *Gv* cluster standard score is plotted against the average of the standard scores from the Visualization and Picture Recognition tests, which compose the *Gv* cluster. If the *Gv* cluster score were simply the average of the Visualization and Picture Recognition tests, all the points in Figure 1 would fall on the diagonal black line. But, the high scores are above the black line and the low scores are below it, meaning that the *Gv* cluster score is more extreme than the average of its parts.

How is this possible? It does not work like this with other types of measurement. If two different measurements indicated that an adult's height was 5 feet 1 inch, it would be absurd to say that the person's true height was 4 feet 11 inches. Why don't composite scores act like other kinds of measures?

[1] The standardization sample data and the *Woodcock-Johnson IV Technical Manual* (McGrew, LaForte, & Schrank, 2014) were consulted many times to generate specific examples and illustrations.

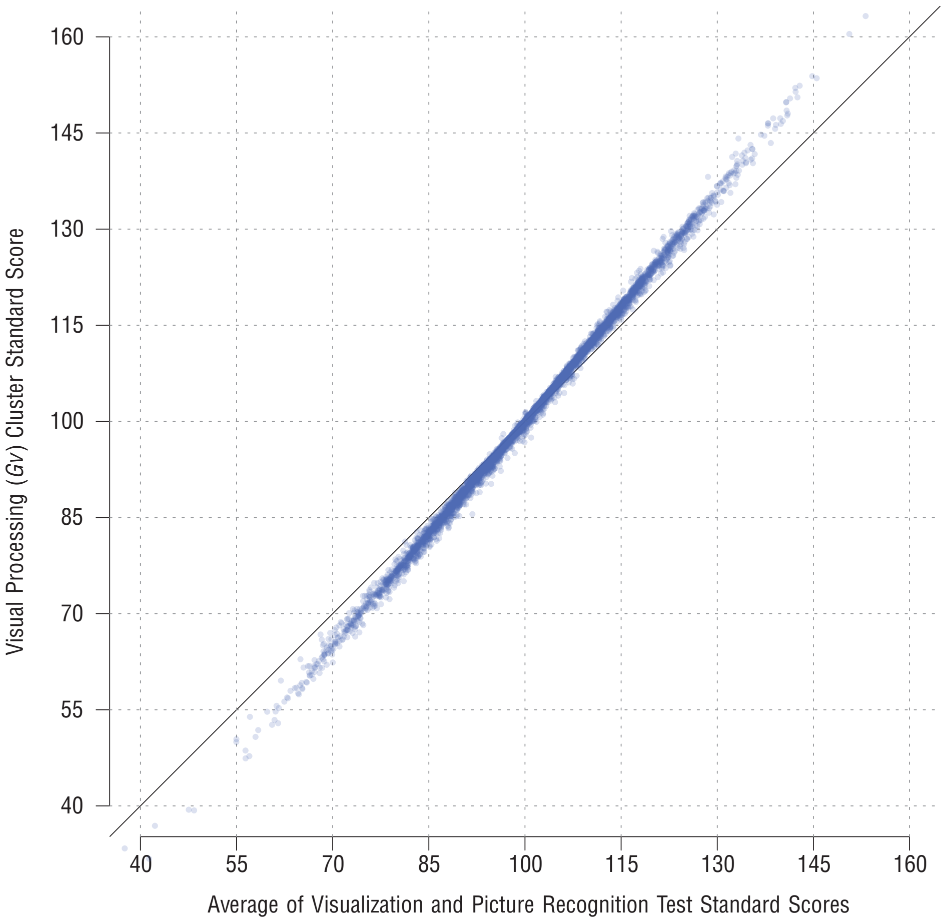


***Figure 1.***
*Relationship between WJ IV Visual Processing (*Gv*) cluster standard scores and the average of the WJ IV* Gv *test standard scores.*

## Composites With Multiple Extreme Scores Are Extremely Strange

The more tests in the cluster, the larger the composite score extremity effect. For example, if a person scored 70 on all seven tests in the WJ IV General Intellectual Ability (GIA) cluster, the cluster score would be around 61 (50 to 65, depending on the examinee's age and pattern of scores). This is even weirder!

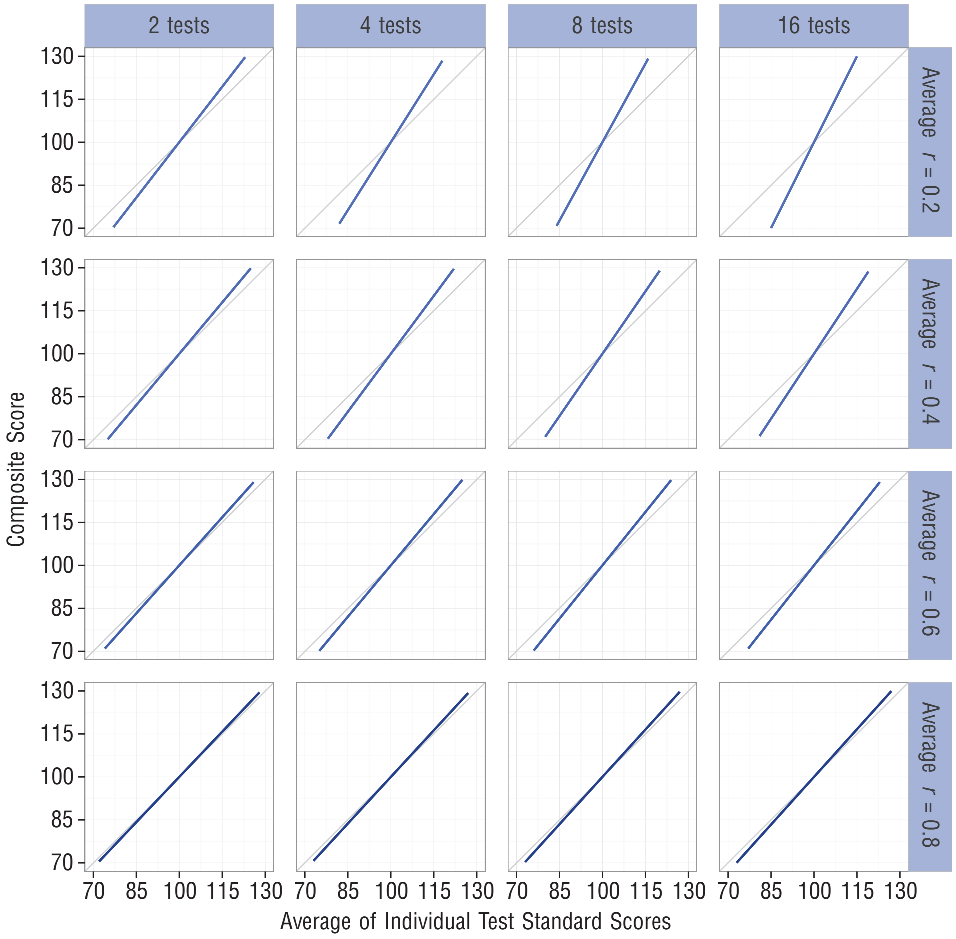
Suppose now that seven different measurements all indicated that an adult's height was 5 feet 1 inch. Imagine saying to that person, "When two different measurements showed you were 5 feet 1 inch, we thought your true height was 4 feet 11 inches. Now that we have seven measurements, all of which show you are 5 feet 1 inch, we have revised our estimate and now believe that your true height is 4 feet 9 inches." This is preposterous! Yet this is precisely the way that composite scores work. Why?

This question arises frequently among users of the WJ IV in particular. When composite scores and their component tests are measured on different scales, as they are on most cognitive test batteries, the odd behavior of composite scores is easy to miss. However, because the WJ IV test and cluster standard scores all have the same mean (*M*)(100) and standard deviation (*SD*)(15), this strange phenomenon is plainly visible. It is especially obvious when the cluster score summarizes many scores, the variables have low correlations, and the scores in question are especially high or low.

The panels in Figure 2 illustrate how the correlations of the test scores influence the composite score extremity effect. In the top row, the correlation is low and the dark lines deviate strongly from the light diagonal lines. In the bottom row, the correlation is high and the dark lines are nearly parallel to the light diagonal lines. That is, the less correlated the tests, the more extreme the composite score. As noted above, the composite score extremity effect is even larger if the composite consists of many tests. Compare the first and last panel columns in Figure 2. When there are only two tests in the composite, the lines deviate a bit from the diagonal, but when the composite summarizes 16 tests, the effect is much stronger, especially when the test correlations are low.

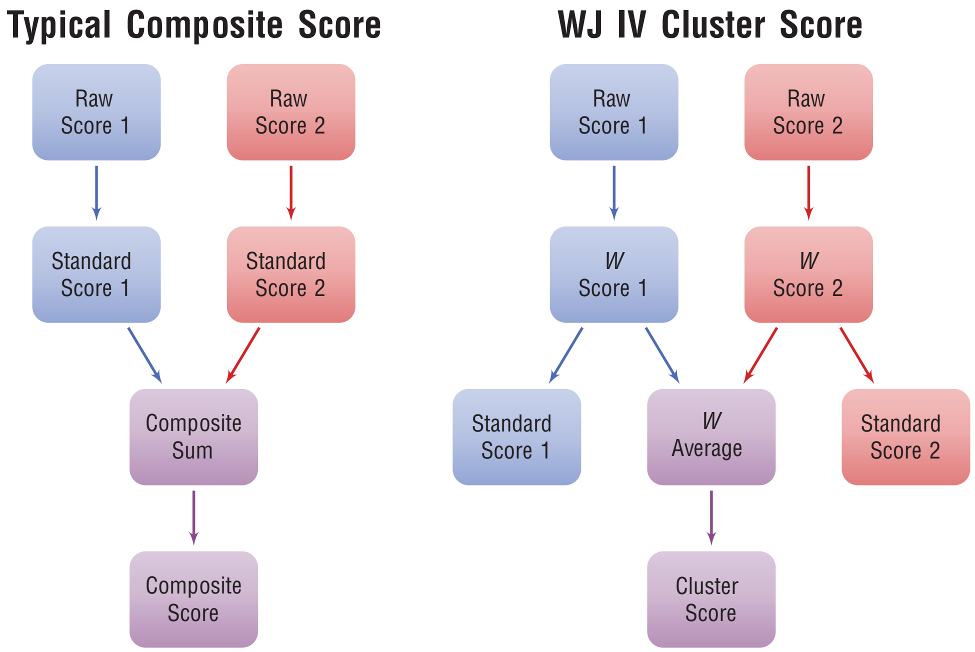
So far, it has been asserted *that* the composite score extremity effect exists; however, it has not yet been explained *how* or *why* it works. Though the composite score extremity effect is deeply counterintuitive at first, once we understand what composite scores are, it all makes sense. Psychological assessment requires a whole new set of intuitions about measurement.

***Figure 2.***
*Relationship between any composite standard score and the average of the contributing test standard scores, by number of tests and correlation between tests.*



As an aside, it should be noted that the procedure for creating WJ IV cluster scores is more complex than it is for creating traditional composite scores, making it difficult to provide a concise mathematical model for their behavior. Figure 3 on page 4 shows how composite scores are typically calculated in other test batteries and how the WJ IV cluster scores are calculated. On most test batteries, raw scores are converted to standard scores directly. The subtest standard scores are added together and then rescaled to create the composite standard score. On the WJ IV, raw scores are first converted to *W* scores, which allow for direct comparison of scores across ages. The *W* scores are then converted to standard scores. The WJ IV cluster scores are created by averaging each test's *W* score. This average is then converted to a cluster standard score. Nevertheless, with respect to the composite score extremity effect, WJ IV cluster scores behave essentially the same way as traditional composite scores do. For the most part, their minor differences will be de-emphasized, and the terms *cluster score* and *composite score* will be used interchangeably.

***Figure 3.*** *Differences between typical composite score calculation and WJ IV cluster score calculation.*

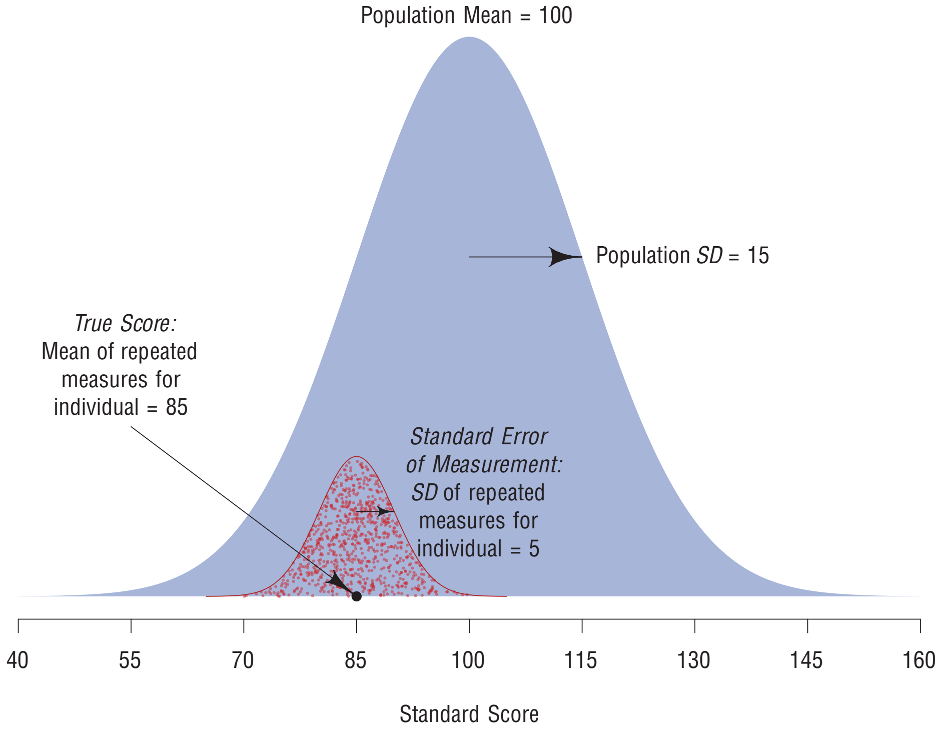


# The Importance of Multiple Measures

If composite scores behave strangely, why do we bother with them? The sad fact is that they are indispensable. Even when implemented flawlessly, psychological measurement typically has so much error that measuring just once is not an option. In one form or another, the use of repeated measures must be employed to reduce measurement error. The most obvious kind of repeated measurement is giving the same test twice.

If we could measure a trait in the same person an infinite number of times with no practice effects—over a wide range of conditions and situations—the average of those scores would be called the *true score*. Figure 4 shows a distribution of scores illustrating what might happen if we could somehow administer the same test to the same person hundreds of times with no fatigue, loss of motivation, or learning between administrations. In this case, the individual's true score is 85, the average of all the red points. If only a single measurement were permitted, we would not have much confidence that the observed score would be close to the true score. The 95% confidence interval ranges from 75 to 95, an interval width of 20 points. However, if several measurements are obtained, there is less chance that the average of the scores would be far from the true score. Why? Measurement errors can either raise or lower a score. When added together, they tend to cancel each other out, which results in a measurement that is closer to the true score.

***Figure 4.***
*Distribution of scores from repeated administration of a test with the same individual.*



### Multiple Items as Repeated Measurements

There are very few single-item psychological measures that are reliable enough for use with individuals. Most well-constructed tests have multiple items. Each item can be thought of as a repeated measurement, and the total score is a summary of all those measurements. Thus, most test scores are a kind of "composite score" in that they are the transformed sum of multiple items. The weird behavior of composite scores operates at this level, too. Answering one difficult item correctly is hard. Answering many such items correctly is even harder. Thus, the standardized total score from many correct answers would be more extreme than the analogous single-item score.

### Composite Scores as Repeated Measurements

In most cases, it is not enough to have multiple test items; multiple test formats increase our confidence that we are measuring the intended construct rather than method variance (i.e., unwanted influences resulting from the test's format). WJ IV cluster scores achieve greater reliability and validity by combining tests that measure the same broad ability but have different test formats. Using diverse tests prevents any one kind of test from having undue influence on the final estimate of the person's ability.

## All Standard Scores Are Strange

The primary reason that composite scores in psychological tests do not behave like other measurements is that they are *standard scores*. Physical quantities, such as distance, mass, and time, are measurements aligned to a particular standard. For example, the meter was at one time equal to the length of platinum bars kept under nearly constant conditions of temperature and humidity. Modern standards for measuring basic physical quantities are even more precise and constant. Standard scores, too, are aligned to a particular standard—two of them, in fact. (They are not called *standard* scores for nothing.) Unfortunately, these standards, the mean and standard deviation of the raw, untransformed scores, differ from test to test and from population to population—they even change over time. We can always transform standard scores to a familiar metric (all scores on the WJ IV have a mean of 100 and a standard deviation of 15). However, the meaning of the standard score is only as stable as its original standards, the raw mean and standard deviation of the normative sample.

The key idea for understanding the strange behavior of standard scores is this: *When test scores with the same standard deviation are averaged, the standard deviation of the average is smaller than the original standard deviation*. For this reason, the averaged score must be transformed so that the composite score has the same mean and standard deviation as its parts. The effect of rescaling the averaged score from a smaller standard deviation to a larger one is to spread the scores out, making them further from the mean and thus more extreme. For example, note how the points in Figure 1 fit into a space that is taller than it is wide.

If we did not rescale our composite scores, every composite would have its own standard deviation, which would interfere with easy interpretation. Thus, the confusion generated by the composite score extremity effect is the price we pay for the convenience of having all of our scores in a standard metric.

# Intuition Pumps for Thinking About Composite Scores

The precise technical explanation of the composite score extremity effect involves a fair amount of formidably difficult mathematics. Fortunately, a formal explanation is not needed to develop a thorough understanding of this topic. Daniel Dennett (1984) used the term *intuition pumps* to describe simplified thought experiments that are "… cunningly designed to focus the reader's attention on 'the important' features, and to deflect the reader from bogging down in hard-to-follow details" (p. 12). So long as the intuition pump is not wildly inaccurate, it is a useful tool. Here are a few ways of thinking about the phenomenon to help retrain your intuition about composite scores.

## Some Strengths and Weaknesses Have Cumulative Effects

We know from everyday experience that having multiple deficits is generally worse than having just one. For example:

- An employee with one area of incompetence is a concern. An employee with many such deficits is a major liability.
- A person guilty of fraud, extortion, arson, and assault is generally more dangerous than someone who has committed just one of those crimes.
- A person with one sprained ankle walks with a limp. A person with *two* sprained ankles probably doesn't walk at all.

Likewise, having multiple advantages is better than having just one. For example,

- A person who has mastered both the piano and the violin is more accomplished than someone who has mastered just one of these instruments.
- A person who earned second place in all 10 events of a decathlon is likely to take first place overall.
- A person with a pleasant personality is a possible date. A person with a pleasant personality, sparkling wit, good looks, and a steady work history is a real catch.

These truths also apply to ability tests. People with a single cognitive or academic deficit will likely find school to be a challenge. With many such deficits, things can become very difficult indeed.

## Having Two Unusual Features Is More Unusual Than Having Just One

It is not unheard of for people to have multiple talents, but usually their talents tend to be clustered in similar domains such as singing, dancing, and acting or math, science, and engineering. As untold numbers of actors with stalled careers can tell you, very few people have sustained success in Hollywood. Scientists seek a different kind of fame, but very few of them can claim to have discovered or invented something truly useful. To be highly successful in either Hollywood or science is fairly unusual. To be successful in both is a very rare thing. For example, Hedy Lamarr (1914–2000) starred in dozens of motion pictures in the 1930s, 40s, and 50s.[2] As a native Austrian with an undisclosed Jewish heritage, Lamarr had been a longtime opponent of the Nazi Party as it rose to power. During World War II, she collaborated on a project designed to send and receive secure communications, hoping that her efforts would help defeat Nazi Germany. Later, her patented inventions became influential in a variety of applications, particularly in

[2] Details about Lamarr's life were drawn from Rhodes (2011).

the field of mobile electronic communications. She was not the greatest actress of her generation, and there were other women whose scientific contributions were more influential. Even so, this unusual combination of extreme talents made her a remarkable individual.

If "general talent" could be summarized with a single number, Lamarr's score would be extreme. To take the average of her acting talent and her scientific accomplishments would not be an adequate summary. Having multiple talents is more unusual than having a single talent. A composite score that summarizes diverse talents must be more extreme than the average of those talents.

Likewise, though it is unusual to have a particular deficit, it is even more unusual to have that deficit and several more. A composite score that summarizes all of these deficits would have to take this comparative rarity into account. It is for this reason that a composite score that consists of many low scores is lower than the average of those scores.

## Extreme Scores Are Doubtful Until They Are Repeatedly Confirmed

If a composite score is a summary of diverse abilities, then it makes sense that a person with multiple deficits should receive a lower score than a person with a single deficit. However, this line of reasoning might not seem to apply if two tests are intended to measure the same ability. If a child scored 70 on two tests measuring the same ability, shouldn't our estimate of this child's ability in what this test measures also be 70? It sure seems so, but in truth our best estimate is actually lower than 70. Cluster scores that measure a single narrow unidimensional ability exhibit exactly the same behavior as cluster scores that measure broad multidimensional abilities. This phenomenon seems bizarre at first, but actually there are everyday experiences in which the cumulative effect of having multiple sources of information is greater than their average, such as with reviews, reputations, and recommendations. As examples:

- When the first review of a movie turns out to be extremely negative, it is a reasonable bet that you would also dislike the film, but probably not as much as the critic did. After all, tastes differ, and what sets off one person might not set off another. However, if all subsequent reviews are equally harsh, the probability increases that you would also have a strong negative reaction to it.
- If one person at your new job says that a particular coworker is very difficult to work with, it is hard to know how seriously to take the warning. However, if everyone else at your new job shares the same opinion, you would do well to be careful around that coworker. A person who is in constant conflict with nearly everyone is, by definition, more difficult to work with than someone who has an ongoing feud with just one person.
- An applicant with a positive letter of recommendation would likely work out well. An applicant with many such recommendations is probably excellent. It is harder to impress multiple people than it is to impress just one.

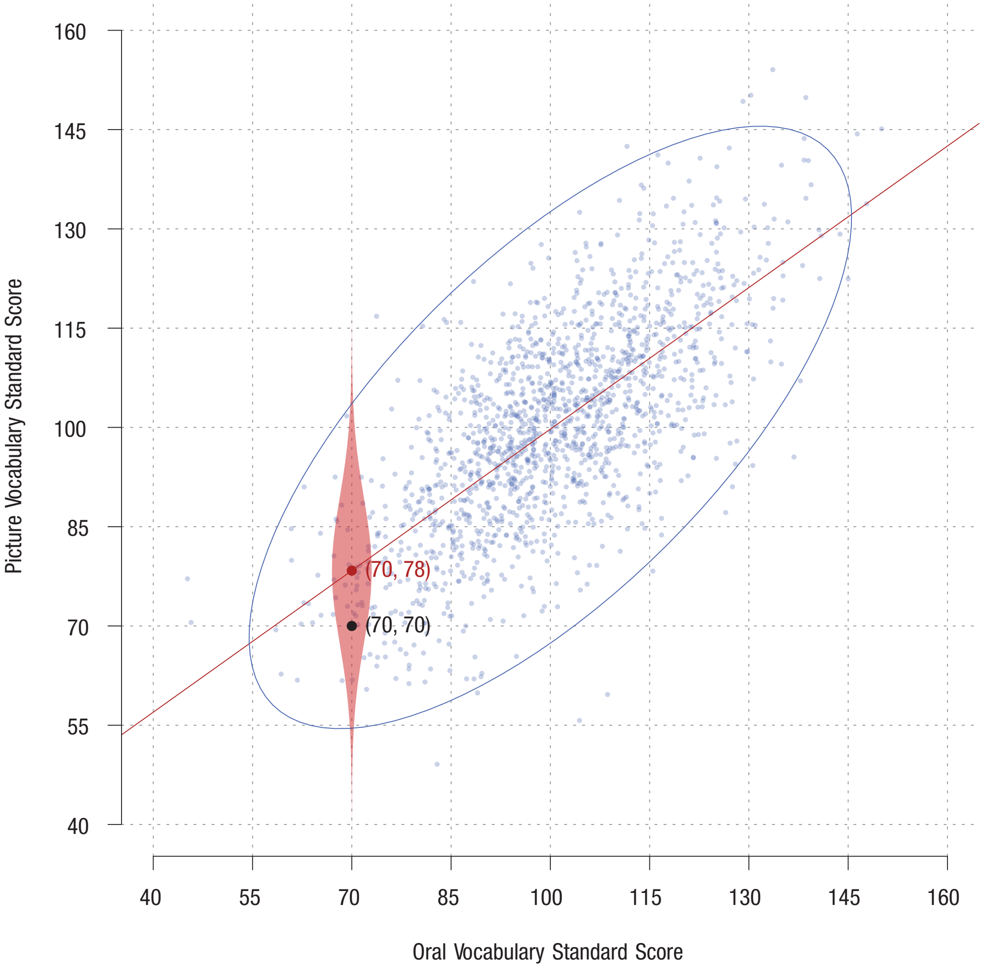
In terms of the WJ IV, the WJ IV COG Oral Vocabulary test and the Picture Vocabulary test of the *Woodcock-Johnson IV Tests of Oral Language* (WJ IV OL; Schrank, Mather, & McGrew, 2014) use different formats to measure the same construct.[3] If a person scores

[3] There are, of course, important differences between these tests that go beyond format. For example, Oral Vocabulary has words that are more abstract and focused on human relations, whereas Picture Vocabulary is understandably more focused on knowledge of physical objects.

very high or very low on both tests, the Vocabulary cluster score (which comprises these two tests) would be even more extreme. Why is this the case? It can be explained in terms of another counterintuitive phenomenon: regression to the mean. In Figure 5, the blue points represent all the individuals ages 9 to 13 in the WJ IV standardization sample. The red regression line represents the predicted Picture Vocabulary score, given a particular Oral Vocabulary score. The red spindlelike shape represents the expected distribution of Picture Vocabulary scores for all the people who score exactly 70 on Oral Vocabulary. Among children who score exactly 70 on Oral Vocabulary, the average score on Picture Vocabulary is 78 (*SD* = 11.1). In this group, a child who scores 70 on Picture Vocabulary is scoring somewhat lower than the predicted score of 78. Thus, the composite score consisting of two subtest scores of 70 must be lower than 70 to account for the fact that people who are equally extreme on both tests are somewhat unusual.

In this case, the composite score is likely to be around 67 or 68. This difference might not seem like much, but with very low scores, differences of a few points can double or halve the associated percentile rank (PR). An index score (*M* = 100, *SD* = 15) of 67 is associated with a percentile rank of 1.4. A score of 70 is 1.6 times as prevalent (PR = 2.3).

***Figure 5.***
*Relationship between Picture Vocabulary and Oral Vocabulary scores in the WJ IV norming sample.*



# Three Technical Explanations of the Composite Score Extremity Effect

The next section is presented for the sake of completeness and fun. If the math is daunting, there is no need to worry—the basic ideas of this paper have already been communicated.

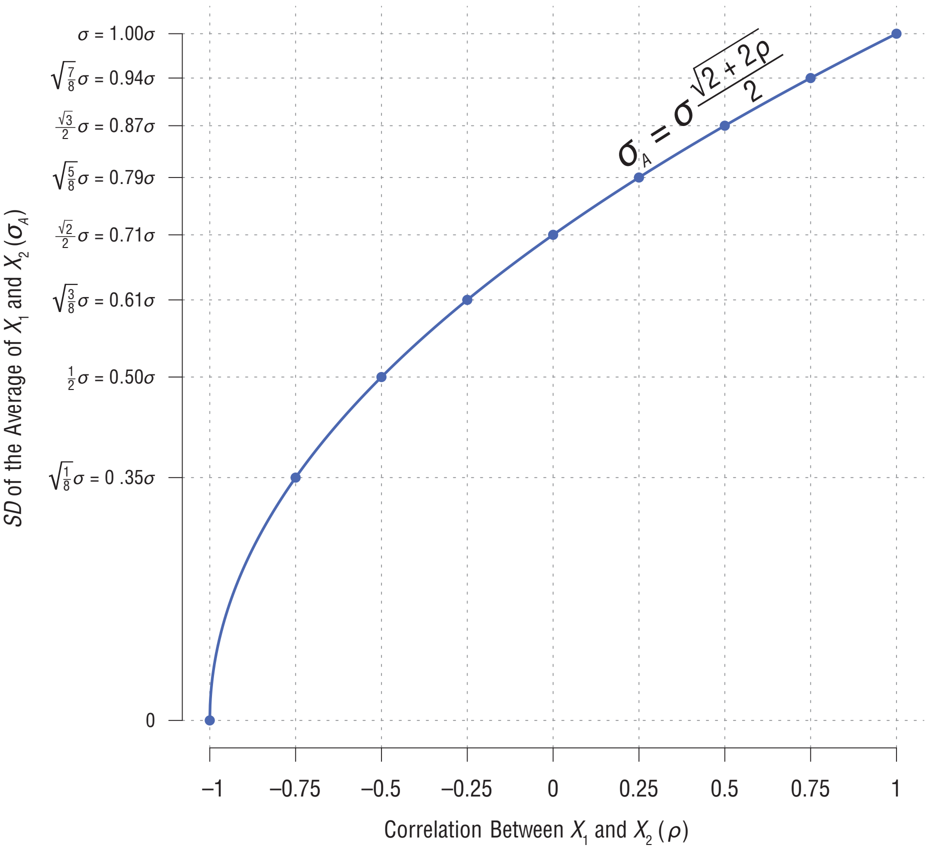
## The Algebraic Explanation

Suppose there are two ability tests, $X_1$ and $X_2$. They both have a mean of $\mu_X$ and a standard deviation of $\sigma_X$. We can call the average of $X_1$ and $X_2$ variable $A$. The mean of $A$ is also $\mu_X$, but the standard deviation of $A$, as seen in Figure 6, depends on $\rho$, the correlation between $X_1$ and $X_2$:

$$\sigma_A = \frac{\sqrt{2 + 2\rho}}{2}\sigma_X \qquad (1)$$

The standard deviation of the *average* of two tests ($\sigma_A$) is less than the standard deviation of the individual tests ($\sigma_X$). The only time that $\sigma_A = \sigma_X$ is when the correlation between the tests is exactly 1, which never happens with real tests.

***Figure 6.***
*The smaller the correlation between two tests, the smaller the* SD *of their average.*



If we wanted to convert $A$ into a proper composite, which we can call $C$, we would first convert $A$ into a $z$ score.

$$z = \frac{A - \mu_X}{\sigma_A} \tag{2}$$

Then we would rescale the $z$ score to have the same mean and standard deviation as $X_1$ and $X_2$.

$$C = z\sigma_X + \mu_X \tag{3}$$

Combining these two formulas, we get the following:

$$C = \frac{\sigma_X}{\sigma_A}(A - \mu_X) + \mu_X \tag{4}$$

There are two features of this formula worth inspecting:

- $A - \mu_X$ tells us how extreme a score is. It is the distance from the mean.
- $\frac{\sigma_X}{\sigma_A}$ is the ratio by which $A - \mu_X$ must be multiplied so that the composite score has the proper standard deviation. Remember that

$$\sigma_X > \sigma_A \tag{5}$$

and therefore

$$\frac{\sigma_X}{\sigma_A} > 1. \tag{6}$$

This means that $C$ will be further from $\mu_X$ than $A$ will be. That is, if $\rho \neq 1$,

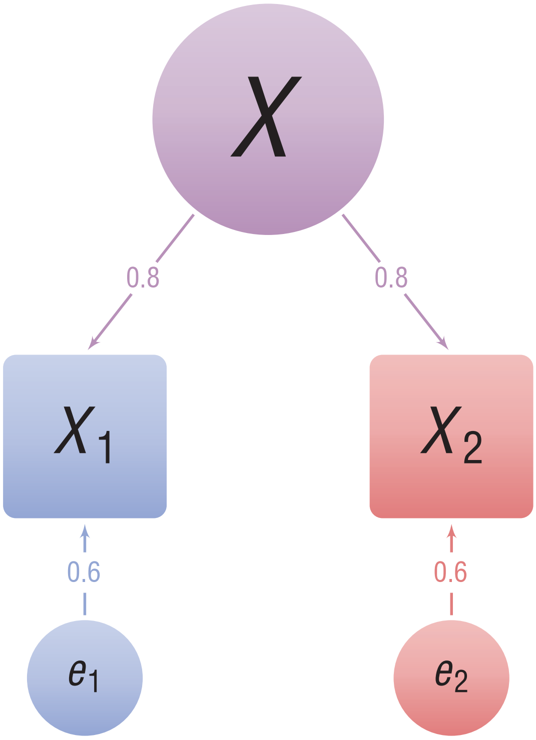
$$|C - \mu_X| > |A - \mu_X|. \tag{7}$$

These formulas can be adapted to show that the same phenomena also occur with weighted composites and with composites that summarize more than two tests. As seen in Figure 3, the procedure for creating WJ IV cluster scores has more steps than the procedure for creating ordinary composite scores, but the basic idea still applies.

## The Latent Trait Explanation

Suppose that $X$ is a latent variable and that $X_1$ and $X_2$ are tests that are designed to measure it, as shown in Figure 7 on page 12. For the sake of convenience, all variables are $z$ scores ($M = 0$, $SD = 1$).

***Figure 7.***
*$X_1$ and $X_2$ as measures of latent variable X.*



*Note*. $X$, $X_1$, $X_2$, $e_1$, and $e_2$ are standard normal.

We will never know the value of $X$ with certainty, but we can estimate it using $X_1$ and $X_2$. Suppose we know that $X_1 = -2$, which corresponds to an index score of 70. $X_1$ and $X$ have a correlation of 0.8. Therefore,

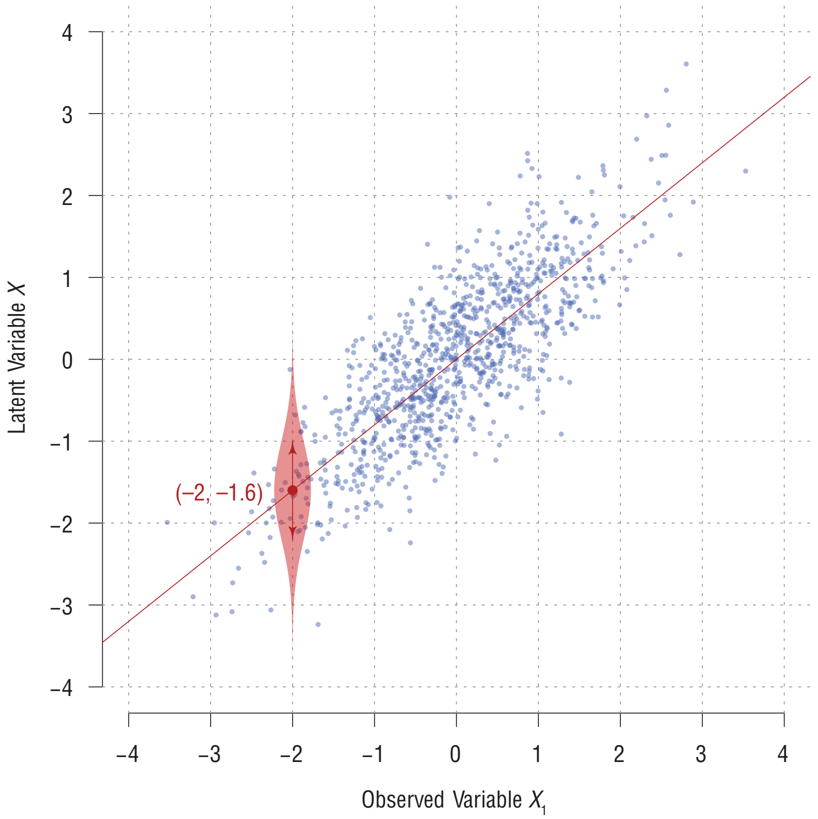
$$\begin{aligned} \hat{X} &= 0.8X_1 \\ &= 0.8 \times -2 \\ &= -1.6. \end{aligned} \tag{8}$$

The standard deviation of $X$ for all people who scored −2 on $X_1$ (i.e., the standard error of the estimate) is

$$\begin{aligned} \sigma_E &= \sigma\sqrt{1-\rho^2} \\ &= \sqrt{1-0.8^2} \\ &= 0.6. \end{aligned} \tag{9}$$

The red regression line in Figure 8 on page 13 shows the predicted value of $X$ for any value of $X_1$. The red spindlelike shape (sometimes referred to as a “violin plot”) shows the expected distribution of scores on $X$ among all people with $X_1 = -2$. So, if $X_1 = -2$, our best guess for the latent variable $X$ is −1.6. Why isn’t our best guess −2? The reason is that $X_1$ correlates with X at $\rho = 0.8$, which is reasonably strong, but far from perfect. When we are using imperfect predictors, we play it safe. Most of the scores are near the mean (0, in this case). So if $X_1 = -2$, we assume that $X$ is below the mean, but odds are that it is not quite as extreme as $X_1$.

***Figure 8.***
*The predicted value of* X *for any value of* $X_1$.

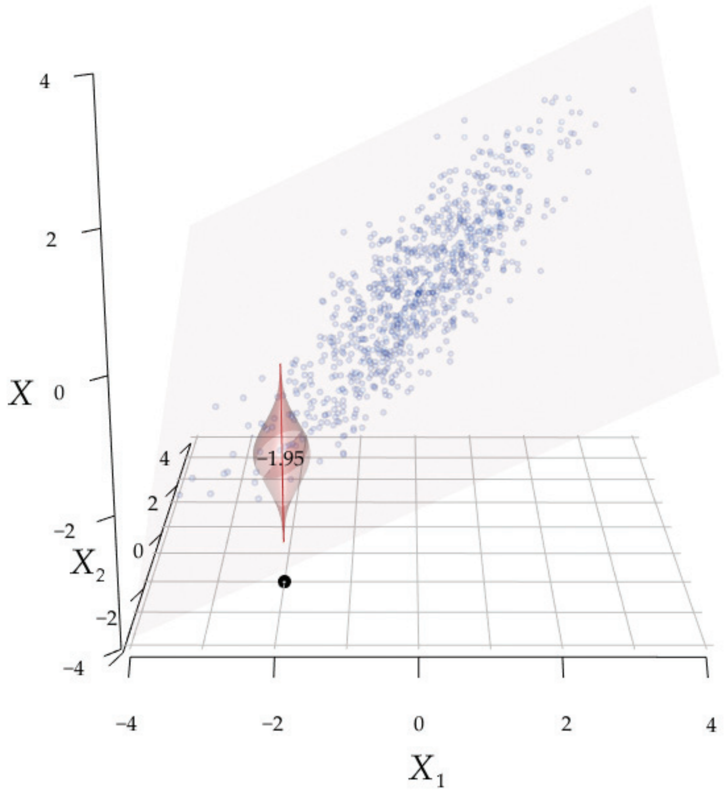


Now suppose we know that $X_1 = -2$ and also that $X_2 = -2$. We can use multiple regression to predict $X$. The correlation between $X_1$ and $X_2$ is the product of their loadings on $X$: $0.8^2 = 0.64$. Applying a bit of matrix algebra (not shown here), we can derive the regression formula for predicting $X$:

$$\begin{aligned}\hat{X} &= b_0 + b_1X_1 + b_2X_2 \\ &= 0 + 0.488X_1 + 0.488X_2 \\ &= (0.488 \times -2) + (0.488 \times -2) \\ &= -1.95. \end{aligned} \tag{10}$$

The regression plane in Figure 9 on page 14 represents the predicted value of latent variable $X$ for each combination of $X_1$ and $X_2$ (see Figure 7). When both $X_1$ and $X_2$ equal −2 (two standard deviations below the mean), the expected distribution of $X$ is represented by the red spindlelike shape, which has a mean of −1.95 and a standard deviation of 0.47. As shown in Figure 9, our best guess for predicting $X$ is now −1.95, with a standard deviation of 0.47. Because we have two test scores, we can predict $X$ more accurately than we did in Figure 8 with just one test score. With greater accuracy, our prediction does not regress to the mean as severely. Indeed, −1.95 is lower than our guess from just one test, −1.6. Analogously, the composite score in this situation is lower than either of the two single scores. Applying the formulas from the previous section, the composite score would be −2.21, which is lower than −2. The correlation between $X$ and $\hat{X}$ (i.e., the Multiple $R$) is 0.883. Interestingly, $-2.21 \times 0.883 = -1.95$. Thus, the composite score regresses to our best estimate of the latent trait.

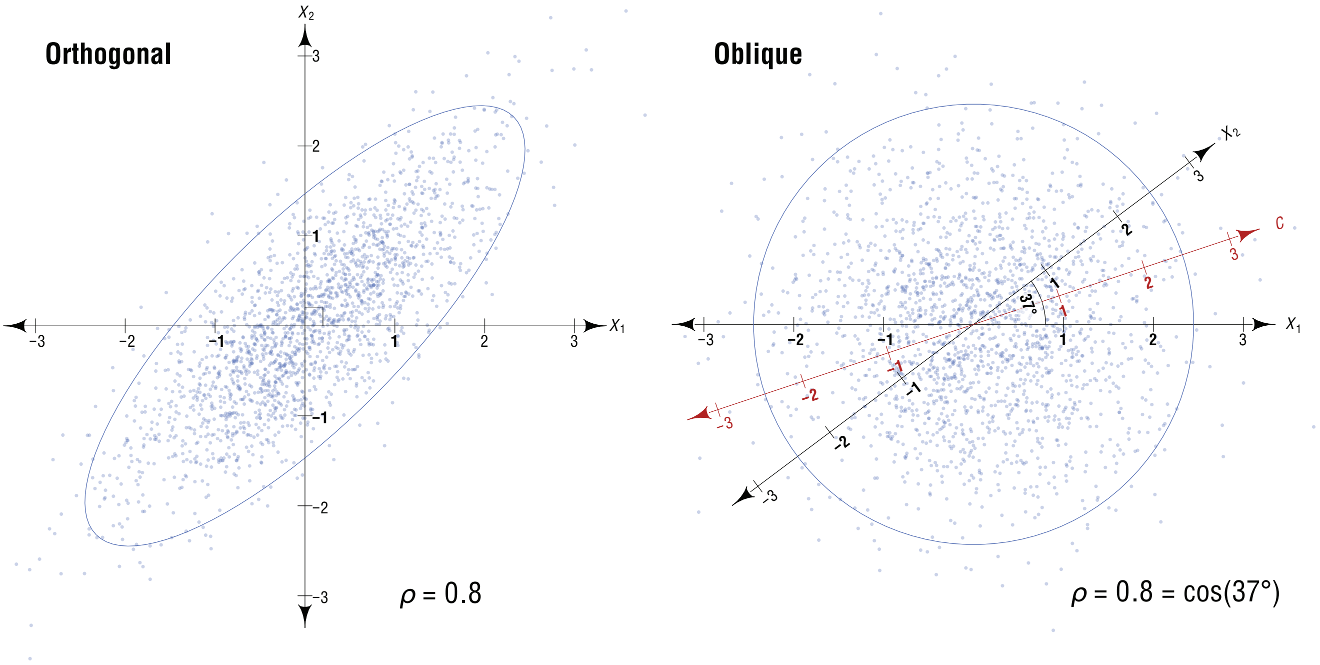
***Figure 9.***
*The predicted value of* X *for each combination of* $X_1$ *and* $X_2$*.*



## The Geometric Explanation

We are used to seeing plots with orthogonal axes (at right angles) such as the plot on the left of Figure 10 on page 15. However, mathematicians have long known that axes can be drawn at any angle. *Oblique axes* are axes that are not at right angles (i.e., not orthogonal). The use of oblique axes is not just a stunt. Oblique axes are regularly used in factor analysis because oblique rotations of the factor loadings make them easier to interpret.

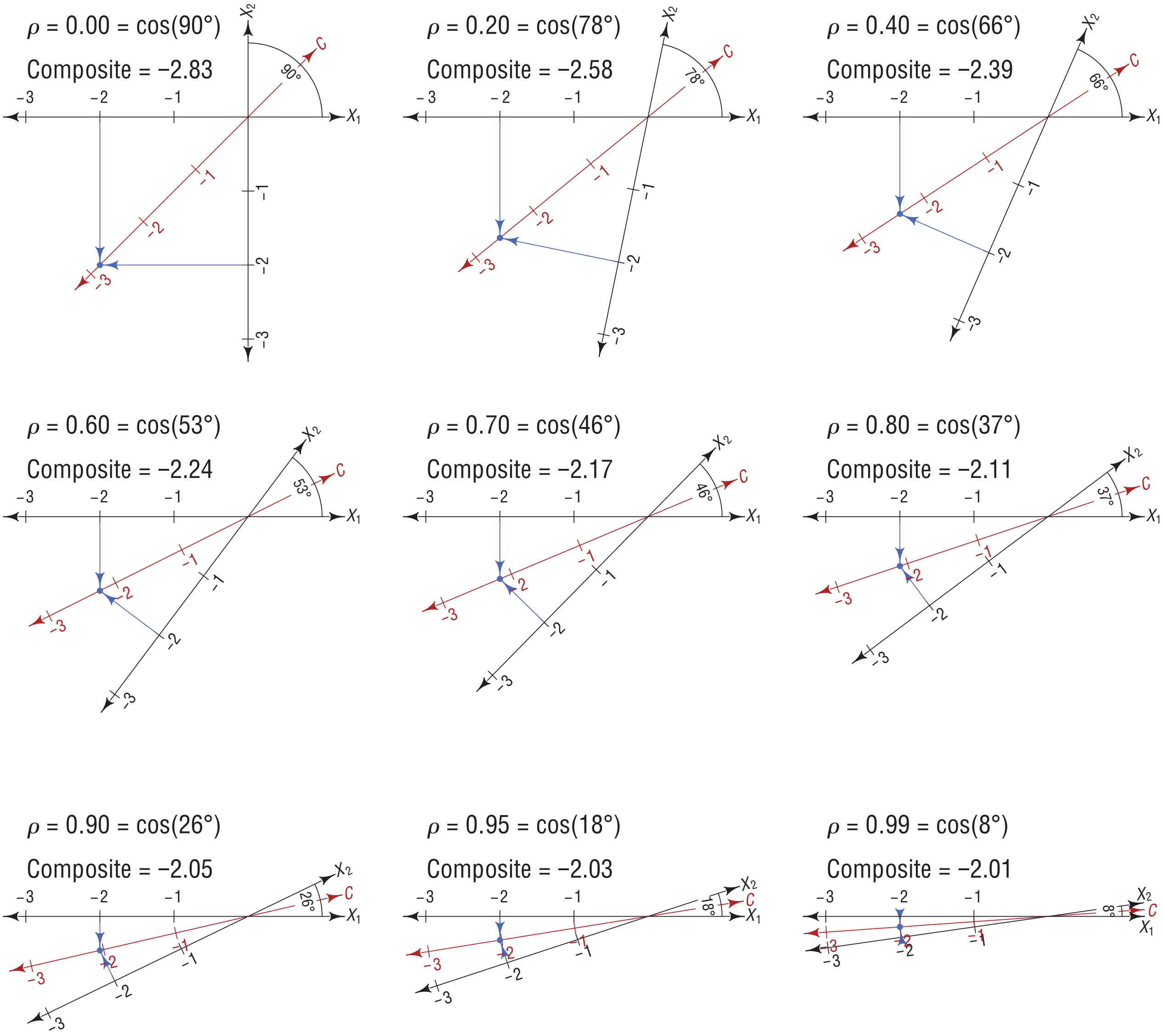
***Figure 10.***
*Correlated data plotted on orthogonal and oblique axes.*



In the case of correlated variables, interesting things happen when the cosine of the axis angle is equal to the correlation coefficient. In the plot on the right of Figure 10, the same data appear to be uncorrelated, but this an illusion. They are just as correlated as they ever were. One advantage of oblique axes is that they allow us to draw a third axis representing the composite score on the same scale as the two tests (e.g., the red axis on the right in Figure 10). Doing so allows us to see the composite score extremity effect as a geometric inevitability.

What happens when we plot the same two scores, $X_1 = -2$ and $X_2 = -2$, on oblique axes? So that the $z$ scores of −2 are comparable to WJ IV scores, we can convert them to index scores of 70. In Figure 11 on page 16, the first panel in the upper left corner represents a composite score consisting of two uncorrelated tests. The blue dot is two standard deviations below the mean for $X_1$ and $X_2$ (index score = 70), but on the red axis, it is about −2.83 standard deviations below the mean (index score = 58). We see that even though the composite score is on the same scale as the tests that compose it, its distance from the origin is greater. In Figure 11, the correlation increases for each panel from left to right and from top to bottom. As the correlation increases, the angle between the axes narrows and the composite score inches closer and closer to 70.

**Figure 11.**
*Composite scores from two tests, $X_1$ and $X_2$, with varying correlations.*



For highly correlated clusters such as the WJ IV Comprehension-Knowledge (*Gc*) cluster ($\rho \approx 0.7$), the difference between the average of the test scores and the cluster score is fairly small, with differences of 2 to 3 points (comparable to the bottom left panel in Figure 11). However, for less correlated clusters, such as the WJ IV Auditory Processing (*Ga*), Long-Term Retrieval (*Glr*), and Visual Processing (*Gv*) clusters ($\rho \approx 0.4$), the difference becomes more noticeable (5 to 7 points). Remember that these are all clusters consisting of only two tests, and the composite score extremity effect is at its weakest. For clusters with more tests, the effect is much stronger. For example, with a seven-test cluster such as the WJ IV General Intellectual Ability (GIA), if the average test score is 70 and the average correlation is around 0.35, the composite score could be around 55, a 15-point difference. This is not a bad thing. With seven tests, the estimate is much more reliable than it would be with only two tests.

## Implications

Practitioners unaware of the composite score extremity effect often worry that something is wrong with the test scoring software when they notice the odd discrepancy between the cluster score and the average of its parts. Hopefully the explanations in this paper provide assurance for those who find the phenomenon unsettling. There is nothing wrong when the composite score is far from the average of its parts. The composite score extremity effect is woven into the mathematical fabric of time and space. It is not a problem we need to solve. Not taking advantage of the additional certainty we have when multiple sources of information are combined would be self-defeating.

The only problems related to the composite score extremity effect that we need to worry about are those that arise when we fail to account for it. For example, some organizations make decisions based on strictly enforced test cut scores. If multiple measures of the relevant construct are available, a composite score should be calculated whenever possible.[4] Simply averaging the scores instead of creating proper composite scores will result in a certain percentage of people on the wrong side of the threshold. In most cases this percentage will be small, but in high-stakes situations (e.g., qualification for special education or death-penalty eligibility decisions), no preventable inaccuracies are acceptable.

## Summary

Composite scores summarize multiple test scores, but they are not averages. The average of multiple scores has a smaller standard deviation than the individual scores do. In order to rescale the composite score so that it has the same standard deviation, the scores are spread out, making them more extreme than the average of the test scores. This makes the composite scores more convenient to interpret, but it has counterintuitive implications. Even if a person scores exactly the same on all of the tests in the composite, the composite score is more extreme. The less correlated the scores, the more extreme the composite score becomes. The effect is also stronger in composite scores that summarize many tests.

There are precise mathematical reasons for this phenomenon, but explanations based on practical experience help make the strange behavior of composite scores less confusing.

- A composite score that summarizes two low scores is lower than the average of the scores because having two weaknesses is worse than having just one. Likewise, a composite score that summarizes two high scores is higher than the average of the two scores because having two advantages is better than having just one.
- It is rare to have one low score. It is even rarer to have multiple low scores. The composite score must be lower than the average of the scores because it must reflect this increased rarity.
- When one test indicates an extreme score, there is reason for interpretive caution. When the extreme performance is confirmed by multiple tests, the confidence in the scores increases.

Although the composite score extremity effect seems counterintuitive, an understanding of its rationale will assist clinicians in accurately interpreting assessment results.

[4] See Schneider (2013) for the computational details and a more complete discussion of these issues.

# References

Dennett, D. C. (1984). *Elbow room: The varieties of free will worth wanting.* Cambridge, MA: MIT Press.

McGrew, K. S., LaForte, E. M., & Schrank, F. A. (2014). Technical Manual. *Woodcock-Johnson IV.* Rolling Meadows, IL: Riverside Publishing.

Rhodes, R. (2011). *Hedy's folly: The life and breakthrough inventions of Hedy Lamarr, the most beautiful woman in the world.* New York, NY: Vintage Books.

Schneider, W. J. (2013). Principles of assessment of aptitude and achievement. In D. Saklofske, C. Reynolds, & V. Schwean (Eds.), *Oxford handbook of psychological assessment of children and adolescents* (pp. 286–330). New York, NY: Oxford University Press.

Schrank, F. A., Mather, N., & McGrew, K. S. (2014). *Woodcock-Johnson IV Tests of Oral Language.* Rolling Meadows, IL: Riverside Publishing.

Schrank, F. A., McGrew, K. S., & Mather, N. (2014). *Woodcock-Johnson IV Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside Publishing.

Houghton
Mifflin
Harcourt™

800.323.9540
www.wj-iv.com