Journal of Pediatric Neuropsychology

Life-and-Death Psychometrics: Generalizable Best Methods for Combining Scores in Intellectual Disability and Other Diagnostic Assessments

W. Joel Schneider, Cecil R. Reynolds, Kevin S. McGrew, and Karen L. Salekin

Online First Publication, March 12, 2026. https://dx.doi.org/10.1037/jpn0000032

CITATION

Schneider, W. J., Reynolds, C. R., McGrew, K. S., & Salekin, K. L. (2026). Life-and-death psychometrics: Generalizable best methods for combining scores in intellectual disability and other diagnostic assessments. *Journal of Pediatric Neuropsychology*. Advance online publication. https://dx.doi.org/10.1037/jpn0000032

Journal of Pediatric Neuropsychology

ISSN: 2199-2673
https://doi.org/10.1037/jpn0000032

Life-and-Death Psychometrics: Generalizable Best Methods for Combining Scores in Intellectual Disability and Other Diagnostic Assessments

W. Joel Schneider[1], Cecil R. Reynolds[2], Kevin S. McGrew[3], and Karen L. Salekin[4]
[1] Department of Psychological Studies in Education, Temple University
[2] Department of Educational Psychology, Texas A&M University
[3] Institute for Applied Psychometrics, Saint Joseph, Minnesota, United States
[4] Department of Psychology, University of Alabama

A diagnosis of intellectual disability is a momentous event that can determine eligibility for special services and supportive sources of income, and in the criminal arena, it can be a matter of life and death. For criminal defendants who might otherwise face capital punishment, it is a matter of life and death. Individuals evaluated for intellectual disability often have been given multiple intelligence tests, sometimes with results falling on both sides of the diagnostic threshold. In all cases, the diagnostic decision must be based on a rigorous examination of the totality of evidence in the context of systematic clinical judgment. When multiple IQ results are relevant and comparable, they can be combined into a properly computed composite score to assist the clinician charged with diagnostic responsibility in determining if Prong 1, deficits in intellectual functioning, of the three-prong criteria necessary for an intellectual disability diagnosis has been met. Best psychometrically grounded methods for these calculations are presented along with a discussion of inappropriate approaches for accurately combining multiple scores. To make these methods accessible to professionals outside the discipline of psychology, all calculations are fully explained in the context of foundational concepts.

Keywords: intellectual disability, intelligence tests, composite scores, multiple IQ scores, capital cases

A diagnosis of intellectual disability (ID) is a life-changing event for the individual and their families. It determines eligibility for special services and supportive sources of income, and in the criminal arena, it can be a matter of life and death. A diagnosis of ID makes one eligible for special education and community support services, as well as supplementary security income (at least in the United States) from the Social Security Administration, family assistance, and, in some states, other additional benefits. Although a diagnosis of ID does not automatically disqualify an individual with ID from obtaining gainful and competitive employment, it is important to note that these individuals face unique challenges in the workplace, including obtaining gainful employment (Carlson et al., 2020; Cavanagh et al., 2021; Gormley, 2015). Individuals with ID can benefit from special hiring practices that allow for a quicker hiring process and provide for reasonable accommodations in many work environments (Lysaght et al., 2012; Madan et al., 2024; Wendelborg et al., 2022). Conversely, preferential treatment may subject such individuals to inaccurate stereotypes and stigma.

Professional standards for the diagnosis of ID have been fluid over decades, as are many other diagnostic criteria, and must be so as our science and understanding of neurodevelopmental disorders progress. Currently, the diagnostic criteria provided in the *Diagnostic and Statistical Manual of Mental Disorders*, fifth edition–Text Revision (*DSM-5-TR*; American Psychiatric Association, 2022), and in *Intellectual Disability: Definition, Diagnosis, Classification, and Systems of Supports*, 12th edition (Schalock et al., 2021), of the American Association on Intellectual and Developmental Disabilities are highly similar and are the most widely adopted ID diagnostic criteria. Briefly stated, an ID diagnosis requires the person under examination to meet three criteria, often referred to as Prongs 1–3. Prong 1 provides that the person displays deficits in general intellectual functions confirmed by clinical examination and the administration of an individually administered, nationally normed, comprehensive standardized intelligence test (Floyd et al., 2021). Prong 2 requires the demonstration of deficits in adaptive functioning that result in the failure to meet current developmental and social standards for personal independence and the establishment of social responsibility. Prong 3 specifies that the onset of the symptoms defined in Prongs 1 and 2 manifests during the developmental period. The use and understanding of the criteria require specialized training and study, especially of the discussions in the relevant diagnostic manuals.

Although all three prongs are essential components of the ID diagnosis, the focus of this article is on the Prong 1 criteria as

Robert J. McCaffrey served as action editor.
W. Joel Schneider https://orcid.org/0000-0002-8393-5316
Cecil R. Reynolds https://orcid.org/0009-0005-3837-1936
Kevin S. McGrew https://orcid.org/0000-0003-4626-7192
Karen L. Salekin https://orcid.org/0000-0001-5308-5771

W. Joel Schneider played a lead role in conceptualization, software, visualization, and writing–original draft. Cecil R. Reynolds played a lead role in writing–review and editing and a supporting role in conceptualization and writing–original draft. Kevin S. McGrew played a supporting role in conceptualization, writing–original draft, and writing–review and editing. Karen L. Salekin played a supporting role in conceptualization, writing–original draft, and writing–review and editing.

Correspondence concerning this article should be addressed to W. Joel Schneider, Department of Psychological Studies in Education, Temple University, 493 Ritter Hall, Philadelphia, PA 19122-6091, United States. Email: schneider@temple.edu

determined by multiple individually administered, nationally normed, comprehensive standardized intelligence tests. In particular, we propose and outline a psychometrically sound procedure where multiple intelligence test scores can be combined into a valid composite score. Similar to the need to consider the measurement error (standard error of measurement [SEM]) inherent in individual IQ measurements, how to calculate a new composite IQ SEM, which will typically be smaller than the individual IQ score SEMs, is presented.

When and how to consider the interpretation of multiple IQs has been addressed in the highest courts of the land (e.g., *Hamm v. Smith*, 2024), as well as in hotly contested legislation. Given that the determination of ID in capital cases has been subjected to such scrutiny and, at times, decisions are based solely on Prong 1 (17%, Blume et al., 2009), it appears that we are dealing with life-and-death psychometrics, a concept unique to capital litigation. It is within this realm that we focus on its application in legal settings. However, the fundamental method can be applied to create composite scores from any acceptable set of scores that constitute repeated assessments of a common construct. Composite scores are used in the assessment of personality traits, career interests, psychopathology, neuropsychological deficits, giftedness, and academic abilities.

When the U.S. Supreme Court ruled that capital punishment for individuals with an ID was unconstitutional (*Atkins v. Virginia*, 2002), setting standards and procedures for identifying ID in capital cases became urgent. Although each state was deemed responsible for setting its own policies for capital punishment eligibility, at times, the Supreme Court has ruled that some states' policies must be revised. For example, in *Hall v. Florida* (2014), the court ruled that the state of Florida must not use a strict "bright-line" cutoff of IQ $\leq$ 70 without considering the inherent measurement error in intelligence test scores. Specifically, when a defendant's IQ is narrowly above the cutoff, the defendant must be allowed to present additional evidence of ID beyond the test scores, in line with professional standards and practices.

Hall v. Florida (2014) was decided by a 5–4 majority. In his dissent, Justice Samuel Alito argued that because Florida allowed Hall to present multiple IQ results, Florida was allowing additional evidence to be presented, and thus, the margin of error was given due consideration:

> There are various ways to account for error in IQ testing. One way is Florida's approach (evaluate multiple test results). Another is to use a mathematical measurement called the "standard error of measurement" or SEM. (p. 15)

It is true that evaluating multiple test results can be completed with psychometrically sound methods that account for and reduce measurement error. Unfortunately, there are intuitively plausible methods for evaluating multiple scores that are technically inadequate and can introduce more error than they correct. In addition, the evaluation of multiple test results is not simply an alternate way of accounting for error that is distinct from an examination of the SEM. A proper analysis of multiple IQs does not end with many (possibly divergent) scores, each with their own SEM. Under appropriate conditions, the scores can be combined into a single, more reliable estimate with a smaller SEM and a narrower confidence interval (Floyd et al., 2021; Schneider, 2013).[1]

In addition to narrowing the confidence interval, properly constructed composite IQ scores have properties that may seem counterintuitive at first. Whereas taking the average of many measurements of physical quantities like height does not result in a biased estimate, taking the simple average of many IQs from the same person does introduce a persistent bias to the score. Fortunately, this bias can be corrected with formulas known and used for many decades. Failing to correct this bias may result in tragic misdiagnoses, particularly in capital cases. If clinicians are to make such life-changing diagnoses, and especially when courts are to make life-and-death decisions informed by estimates of intelligence, the test score estimates should be based on the best methods available.

The fundamental methods we present are not new. They rest on principles known since the mid-19th century (Bienaymé, 1853). The methods are not merely opinion but are based on strong mathematical proofs (e.g., see Horst, 1936; Wilks, 1938) that can be reproduced and verified by anyone with graduate-level training in statistics (Ley, 1972). Though elsewhere the methods have been refined, extended, and generalized beyond their original applications, there is no controversy about whether the methods themselves are correct when they are properly applied in the right contexts. The methods are used implicitly or explicitly in nearly all psychological assessment instruments. That is, when scales are created by combining multiple test items or multiple subscales, by convention, the resulting variable is rescaled to have a standard score with a conveniently round mean and standard deviation (e.g., 100 and 15, respectively). If there is a new idea here, it is this: *When multiple, comparable IQs are available, they should be combined using the same methods that we combine every other kind of psychological test scores in every other context*. There is nothing special about IQ that makes it the sole exception as to how multiple scores should be combined.

Although the methods we present in this article have been presented in abbreviated form in book chapters about ID assessment (Floyd et al., 2021) and intelligence testing in general (Schneider, 2013), we cover the topic in greater depth. Although the methods are quite general—they apply to any construct that is measured with more than one instrument with interval scaling—we tailor our recommendations to emphasize forensic assessments in the context of cases in which there is the possibility of capital punishment.

Because we hope this article will be useful to professionals from disciplines outside of psychology, we take time to define technical terms that scholars typically omit when addressing insiders (e.g., Table A1), and give an expanded introduction to technical concepts (e.g., Appendix B). This article will focus primarily on the following:

1. Common intuitive but incorrect methods
2. The composite score extremity effect (CSEE)
3. Measurement error and the related concepts of true scores, reliability, and confidence intervals
4. ID criteria
5. Composite scores and the computation
6. Recommendations for creating composite IQs

[1] The professional consensus is that a 95% confidence interval—approximately ±2 SEM—should be used in ID diagnoses, including in capital cases (Floyd et al., 2021; McGrew, 2015; Watson, 2015).

Common Intuitive Methods That Are Incorrect: Averaging, the Median, and Best Score Approaches

Psychological assessments often generate such a large number of scores that they need to be organized and summarized so that a useful and coherent case conceptualization is possible. Although there are a variety of methods for interpreting multiple scores measuring the same psychological construct, not all of them are suitable for high-stakes decisions and final, formal communication of assessment results. In our work, we have seen others apply intuitively appealing multiple-score "averaging" methods that produce inaccurate and, at times, inherently biased results, especially from IQ measures.

Simple, convenient procedures might be accurate enough for rapid, informal estimates at an early stage of the interpretive process (e.g., when a clinician is deciding whether an inconclusive finding needs follow-up testing), but the final presentation of assessment results must be based on scores computed using the best available methods, convenient or not. When multiple measures can be appropriately summarized with a single score, we recommend creating a new composite score. The computation of such a score may be a little less convenient than other methods, but the interpretation of a composite score is far simpler and most closely mimics how IQ test scores are created by the makers of the tests themselves. The computational advantages of other methods are more than offset by the inaccuracies and/or interpretive difficulties they introduce.

The most common suboptimal method of combining scores that we have observed is the calculation of the arithmetic average, where all scores are summed and divided by the number of scores. It is also common for clinicians to take the median score, especially when one score is a clear outlier. We have also seen clinicians and forensic professionals designate the highest obtained IQ test score as the best estimate of general intellectual functioning based on the incorrect logic that "one cannot be smarter than they are—and the highest score then must be the most accurate."[2] These approaches do not produce sufficiently accurate results in psychological measurement for multiple reasons. As often occurs, however, statistical issues can be counterintuitive and require a more detailed explanation for understanding the phenomenon under review.

Because all well-designed measures of IQ are strongly and positively intercorrelated (Breit et al., 2024), when an IQ is well below 70 or well above it, a second IQ test score will usually also fall on the same side of the diagnostic threshold. However, because IQ test scores are imperfectly correlated, when individuals with low cognitive abilities near the ID diagnostic threshold are given multiple tests, there is a substantial probability that at least some scores will fall on opposite sides of the threshold (Schneider et al., 2024).

When faced with scores that are not on the same side of a diagnostic threshold, stakeholders (e.g., clinicians, lawyers, and triers-of-fact) may default to the application of statistical measures of central tendency and dispersion to decide whether the diagnostic criteria have been met. Taking the arithmetic mean value as the best estimate has intuitive appeal because the mean is normally an excellent measure of central tendency. Likewise, taking the median value might normally be a reasonable alternative if there are strong outliers of dubious accuracy, such as when one of the scores is implausibly higher or implausibly lower than the other scores (Floyd et al., 2021).

Unfortunately, both the mean and the median introduce a known bias that is routinely corrected in most other contexts and that ought to be corrected in this one as well. The standard deviation of a mean or median of a set of scores will be smaller than the standard deviation of 15 that we expect to see in an IQ. A score with a smaller standard deviation has a different interpretation. Thus, the new computed variable can be rescaled to have a standard deviation of 15 so that it can be interpreted as an IQ. The consequences of this correction can be counterintuitive when first encountered. Thus, we will introduce the topic by way of a useful analogy before presenting it formally and documenting its effect on obtained IQs.

How IQ Is Less Like a Ruler and More Like the Decathlon

When faced with counterintuitive phenomena, Dennett (2013) advocated the use of *intuition pumps*—analogies that readily communicate a point, even if the comparison is not technically perfect on every dimension. If we measured someone's height repeatedly, and each result was close to 162 cm (5 ft, 4 in.), the best estimate of the person's height is the average of these measurements: 162 cm. Imagine how strange it would be if we claimed that the best estimate of the person's height was 158 cm, despite the fact that each measurement was well above 158 cm. This claim is clearly absurd. However, in psychological measurement, seemingly similar results happen all the time, and they are correct. How is this possible?

Averaging measurements is not always a good idea because not all variables that look like numbers behave the way we normally expect numbers to behave. If a friend has multiple telephone numbers, the average of those numbers would not be a more accurate way to reach that friend. Numbers used for identification (e.g., telephone numbers, social security numbers, and postal codes) cannot be added, subtracted, multiplied, or divided, much less averaged in any meaningful way—these numbers are on a *nominal* scale of measurement (i.e., the categorization of information into groups or labels without any inherent order) and fail to maintain meaning in the face of arithmetic manipulation.

Ordinal numbers are used to rank some variable of interest (1st, 2nd, 3rd, and so forth). Although it is possible to average ordinal numbers, the interpretation of such averages is not always straightforward. For example, in the 2024 Summer Olympics, Norwegian athlete Markus Rooth completed all 10 events in the Men's Decathlon. Figure 1 shows his final rank in each event.

Should we feel bad for Marcus Rooth? Better luck next time? He did not place first in any event! Out of 10 events, his average rank was 7.4. Did he finish in 7th or 8th place? No, he did not. This performance earned him the Gold Medal and, by tradition, confers upon him the title of "World's Greatest Athlete!"

Why is Rooth's average rank so different from his overall rank? Is this some kind of fluke? In an extraordinary performance in the 2021 Tokyo Olympics, Canadian athlete Damian Warner won Gold by setting the Olympic record for the Men's Decathlon event. He had an average rank of "only" 4.6 and placed first in "only" three events.

[2] If IQ tests were perfect and malingering were the only reason a person would score lower than their potential, this reasoning would hold. In reality, there are many reasons a person's score can fluctuate up and down because of chance factors. Taking only the highest score as the best estimate is wrong for the same reasons it is wrong to calculate a batting average using only a baseball player's best performance. Such a procedure would upwardly bias IQ estimates, making them less accurate predictors of everyday performance in academic and occupational settings.

The average rank is clearly a biased way to estimate overall performance.

When decathlon officials tally up the points to determine the winner, they do not average the event rank. That would be silly, and for an obvious reason: No one would ever win. Although it is theoretically possible to have an average rank of 1 by winning every event, not even Jim Thorpe did so when he dominated the field in the 1912 games in Stockholm (average rank of 2.4).

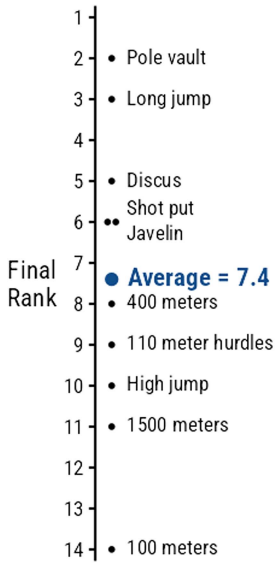
What does this discussion of decathlon performances have to do with IQ? Suppose a person scored exactly 75 on seven intelligence tests. The person's IQ is probably 75, right? Actually, no. If those seven intelligence tests were the seven subtests of the Wechsler Adult Intelligence Scale, Fifth Edition (WAIS-5; Wechsler, 2024)—a subtest scaled score of 5 equals 75 on the IQ metric—then the Full Scale IQ would be 69. Readers with access to the WAIS-5 manual can verify for themselves that this is true. How is it possible—how does it even make sense—that someone without a single subtest score below 75 could be said to have an IQ of 69? Counterintuitive though it may be, roughly for the same reason that Markus Rooth won a Gold Medal in 2024, this is how intelligence scores and most other psychological measurements work.

Physical quantities like distance and mass are measured on what is called a *ratio* scale. On a ratio scale, a zero is a true zero, indicating an absence of what is measured (e.g., 0 g indicates no mass). Ratio measurements, when averaged, remain on the same scale as the original measurements.

Psychological and educational measurements typically require an *interval* scale, which lacks a true zero. If an interval scale happens to include zero, the zero might have a special meaning (e.g., 0 on a *z*-score corresponds to the population mean), but it does not indicate the absence of what is measured. IQ, like most measurements of individual differences, is an interval measurement that has been standardized to have a particular mean and standard deviation. The arithmetic average of multiple standard scores retains the same mean as the original scores, but the average has a smaller standard deviation, which alters its interpretation (in the same way that converting inches to centimeters alters the interpretation of the numbers). To give the average of multiple standard scores the same interpretation as the original scores, the averaged score must be rescaled to have the same standard deviation. The necessity of rescaling the average has surprising implications.

Figure 1

Marcus Rooth's Decathlon Performance at the Paris 2024 Olympics



The CSEE

Although it is rare for someone to have one unusual performance, it is rarer to have multiple unusual performances. Although Markus Rooth did not have the best performance in any single event in the 2024 Olympics, he did unusually well across 10 events, and thus, his relative performance cumulatively exceeded that of everyone else in the competition.

A WAIS-5 subtest score of 5 (i.e., 75 in the IQ metric) is unusually low, at the 5th percentile. To score at the 5th percentile or lower consistently across seven subtests is more unusual than scoring at the 5th percentile or lower on a single test. Because only 2% of people score this low consistently, obtaining seven subtest scores equivalent to 75 results in a WAIS-5 Full Scale IQ of 69, a full 6 points lower than 75. The effect is not specific to the WAIS-5 nor to the Wechsler tests. If a person scored the equivalent of 80 on the intelligence subtests from the Reynolds Intellectual Assessment Scales–Second Edition Normative Update (Reynolds & Kamphaus, 2026), the Composite Intelligence Index (i.e., IQ) would be 75, not 80.

This phenomenon, dubbed the *CSEE* (Schneider, 2016), is not specific to IQ either. It happens whenever standardized scores based on deviations from population norms are combined into composite scores that have the same population means and standard deviations as the tests they summarize. Composite scores are a feature of most educational and psychological test batteries. This effect is also the reason test developers do not use a simple linear conversion of an average to compute a composite score—a sum of scores is completely rescaled to create a composite that is an accurate representation of standing on the latent variable in question relative to others in the chosen reference or normative group (also see Reynolds et al., 2021). An average could also be rescaled—though this would introduce an extra and unnecessary step in the scaling process—but whatever representation of the cumulation of the part scores is used, it must be rescaled. Although the examples presented here have been associated with low scores, it should be noted that the effect works in both directions—obtaining consistently high subtest scores is associated with an even higher Full Scale IQ.

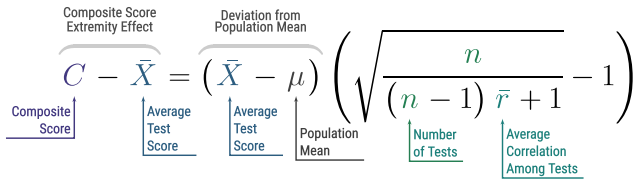
A composite score (C) is more extreme than the average of the tests ($\bar{X}$) used to compute the composite. The effect can be small and subtle, or it can be quite large. The difference between the composite score and the average of the part scores ($C - \bar{X}$)becomes larger when:

1. The scores are, on average, increasingly extreme compared to the population mean (i.e., $|\bar{X} - \mu|$ is large).
2. The correlations among the scores, on average, are increasingly low (i.e., $\bar{r}$ is near zero).
3. The number of scores (n) is increasingly large.

The precise relationship between these three factors and the size of the CSEE is specified by the equation in Figure 2. These relationships are illustrated in Figure 3. If the tests were perfectly

Figure 2

Factors Contributing to the Composite Score Extremity Effect



Note. All tests are assumed to have the same standard deviation.

correlated, there would be no CSEE, and the scores would fall on the black lines in Figure 3 (i.e., the composite score would equal the averaged score). As the correlations weaken, the composite score deviates more strongly from the black line. This effect is stronger on the right panel of Figure 3 than on the left panel because the right panel has more tests. The main point for the clinician to understand is that the CSEE is strongest when a person has extremely high or extremely low scores on many poorly correlated measures. The CSEE is weaker when there are only two highly correlated tests, and the effect is completely absent when a person's average performance exactly matches the population mean.

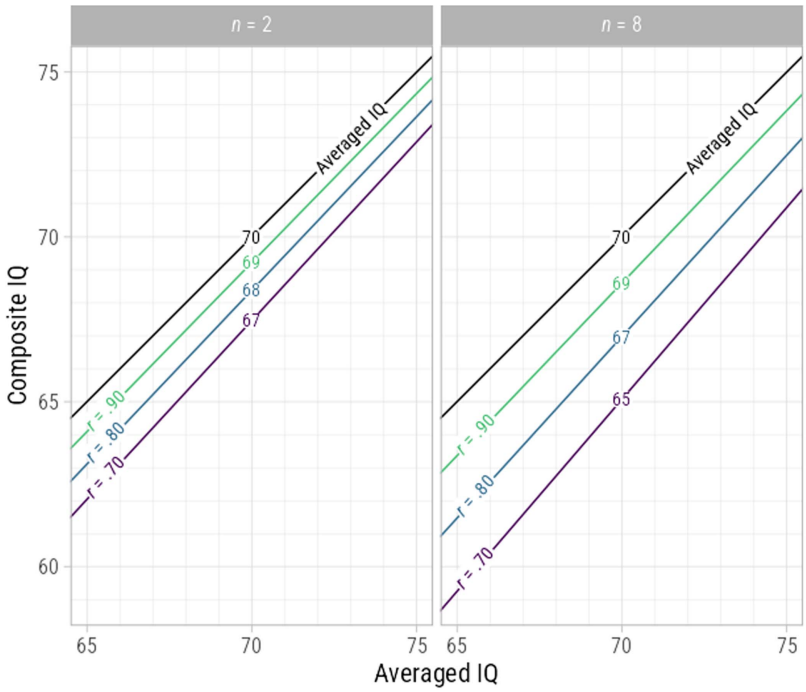
Why does the CSEE happen with psychological test scores and not with, for example, height measured in centimeters? If we were silly enough to measure height with standard scores instead of centimeters, the CSEE would be relevant for height, too. Fortunately, measuring height in centimeters works quite well because we can compare distances to known physical standards. With most psychological variables, there is no mental meter stick in Paris to which test scores can be compared. Lacking an external standard, we create standard scores by comparing psychological variables to two statistical standards, the population mean and the population standard deviation. For example, an IQ of 70 is interpreted in the context of a population mean of 100 and a standard deviation of 15. Thus, an IQ of 70 is 2 standard deviations below the mean ($70 = 100 - 2 \times 15$).

The average of multiple height measurements is still interpreted as having the same metric as the height measurements (i.e., centimeters). When standard scores are averaged, the average still has the same population mean, but its standard deviation is smaller than the standard deviation of the individual scores. For example, suppose we administered five different IQ batteries to everyone, each with a mean of 100 and a standard deviation of 15. The averaged IQ would have a population mean of 100, but the standard deviation of the average IQ would be less than 15, depending on the strength of the correlations among the tests. Not having the same standard deviation, the interpretation of the averaged score is not the same as the interpretation of the individual scores. To put the averaged score on the same metric as the individuals' scores, a correction needs to be applied, the same correction that is applied to composite scores.

This is precisely the process by which test developers create normative scores or reference tables for composite scores. Subtest-level scores are rescaled composites of summed item-level scores, and IQs are rescaled composites of a sum of subtest-level (the part scores) scores. To combine scores across different tests, a procedure is required that allows the new composite score to reflect the actual distribution of the new composite and that mimics how normative scores are derived, and the methods for this calculation are known and given later, below.

Figure 3

Composite IQ as a Function of Averaged IQ, Number of IQ Test Scores (n), and Average Correlation Among the Scores (r)



IQ and Measurement Error

In everyday activities—baking, home repairs, and art projects—we can apply the carpenter's adage to "measure twice, cut once." Most often the measurements will be the same, and we can proceed with confidence. If two measurements differ, a third measurement will likely tell us which of the two measurements is correct.

Some measurements are inaccurate because of preventable mistakes, some because of limitations of the measurement instrument. However, perfect measurement is an ideal that is not obtainable. A hard-won insight from physics is that measurements of continuous quantities like distance, weight, and temperature are not truly fixed but fluctuate from measurement to measurement (Taylor, 2022, pp. 4–5)—not necessarily because there is something wrong with the measurements, but because the underlying phenomena themselves are in flux and the act of measurement can disturb the very phenomena being measured (Heisenberg, 1927). Fluctuations from measurement to measurement, whether they are due to mistakes, instrument limitations, or inherent instability of the phenomena, are called *measurement error*.

To the common errors of measurement in physically observable variables, psychological testing adds errors related to ability domain sampling—we cannot ask every possible question or present every possible problem that is relevant to evaluating or measuring intelligence (or other psychological constructs), and so we sample the ability domain of potential questions/problems.

Usually, the fluctuations of physical quantities are so small that they do not matter for everyday projects. By contrast, measurement errors in psychological assessment are large enough that there can be no assurance that any particular measurement is the correct one. We can, with repeated measurements, reduce measurement error to the point that we can specify a usefully narrow range for where a quantity is plausibly located.

Most psychological measurement devices are not single measurements but consist of multiple test items that are added to produce a total score. IQ tests consist of smaller tests (i.e., *subtests*), each with multiple items. Each test item can be viewed as a single measurement. A core psychometric principle is that when an increasing number of items are aggregated into a single test score, measurement error is further reduced—provided the items are substantially correlated with each other (Brown, 1910; Paunonen, 1984; Spearman, 1910). Even with many test items, however, psychological measures have much more measurement error than do measures of physical quantities such as mass, distance, and temperature. Thus, we expect nontrivial fluctuations every time we measure a psychological state or trait. Indeed, this is why the inherent error of a test must be considered when a psychological test is used to make an important legal decision (e.g., *Hall v. Florida*, 2014).

Biases Introduced by Errors of Domain Sampling

The domain of intelligence comprises a large and diverse collection of interrelated abilities (Carroll, 1993). To the extent we make errors (and we do) in sampling the domain adequately, we introduce another source of measurement error. We also have different and still useful definitions of intelligence without perfect consensus in the field. The aspects unique to each definition arise from theories of intelligence, which, despite being similar, are not the same. Because intelligence tests are based on different theories of intelligence, the emphasis placed on different areas of problem solving varies across different tests of intelligence (McGrew, 2015). Unlike errors due to random fluctuations, errors of domain sampling introduce persistent measurement errors (i.e., measurement biases). Unfortunately, this variability can enhance perceived errors of domain sampling. This is a key reason routinely taking the highest score from a set of obtained IQs is poor practice—an idiosyncratic strength on a unique and narrow component of a particular test will lead to the overestimation of a person's overall ability.

Errors of domain sampling across tests and differences in construct definition and relative emphasis all affect the content of the measures used. As we add more measures, the likelihood of chance factors related to differences in domain sampling creating nearly random fluctuations across measuring devices increases, and the highest obtained score most likely reflects taking advantage of such chance or random variations to produce a speciously high estimate of intellectual functioning. Added to this issue, the distribution of the highest score taken from a set of correlated scores has a higher mean and smaller standard deviation compared to IQ. When there are more than two correlated scores to select from, the means, standard deviations, and probability functions for extreme score distributions are not obtainable with closed-form equations and must be estimated via advanced methods (Botev et al., 2015), making their calculation and interpretation anything but straightforward. In short, substituting the highest score from a group of scores as opposed to calculating a true composite does nothing to solve the problem of creating an accurate score distribution for comparison.

True Scores and Construct Scores

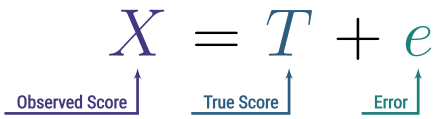
If psychological measurements simply will not hold still, how can one use psychological test results to make high-stakes decisions? It is possible to reduce measurement error with repeated measurement. In classical test theory (Lord & Novick, 1968), we can think of each measurement X as the sum of a *true score* and a *measurement error* (Furr, 2022), as seen in the equation in Figure 4.

The *true score* is defined as the hypothetical average score a person would obtain if a particular test could be given repeatedly without *carryover effects* (Lord & Novick, 1968, pp. 27–28). In the absence of carryover effects (i.e., the influence of previous measurements on subsequent measurements), participants never tire, lose motivation, or learn how to do the task better via practice. Imagine we could somehow rewind time and give a particular test an infinite number of times under an infinite number of plausible conditions. In this thought experiment, the true score would be the average of all possible score fluctuations on that test. Ideally, the fluctuations are small enough that the measurements fall within a usefully narrow range. For example, we may not know precisely how well a person retains information, but the overall score on a series of memory tests can tell us whether a person needs support and treatment for memory problems.

The *true score*, despite its name, is not THE TRUTH in any final sense. A test is designed to measure a theoretical entity or "construct" (e.g., intelligence, anxiety, or reading comprehension). No test measures its intended construct perfectly and in its entirety. Because different intelligence tests emphasize different aspects of intelligence, a person's true score on one test might differ somewhat from the person's true score on another test. Any persistent flaws in a test will be passed along to the true score for that test. When people

Figure 4

Observed Scores Are the Sum of True Scores and Errors



incorrectly assume that a true score perfectly reflects its intended theoretical construct, what they have in mind is a hypothetical idea called a *construct score* (Borsboom & Mellenbergh, 2002), which precedes and is wholly independent of any measurement.[3] When multiple IQ tests are given to a person, the intent is to approximate the construct score. That is, we care about a particular measurement only to the extent that it helps us specify the likely range in which the construct score is located. The true score for the combination of many scores from well-validated IQ batteries provides the best approximation of a construct score for overall intelligence.

Measuring Reliability With Reliability Coefficients

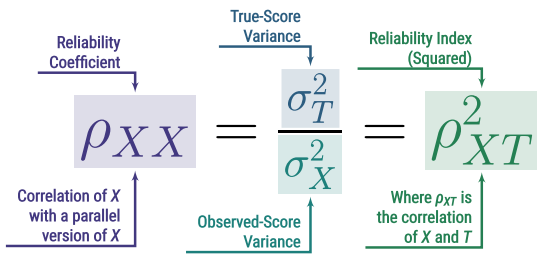
If we could rewind time and measure abilities repeatedly without carryover effects, the scores will not be the same every time, but ideally the scores remain fairly consistent (presuming the target trait is stable). The degree of consistency can be measured with a reliability coefficient. In classical test theory, reliability is a ratio of true score variance to observed score variance (where *variance* is the standard deviation squared).

In Figure 5, the reliability coefficient is symbolized as ρ_{XX} because ρ is the symbol for the correlation coefficient in a population, and the reliability coefficient can also be viewed as the correlation of a test X with a parallel version of X, or, alternately, the short-term retest correlation of X given twice if there are no carryover effects. If there are no parallel versions of a test, and it is only given once, it is possible to estimate a test score's reliability coefficient from the relationships among the items that make up the score. In essence, each item of the test can be thought of as an alternate version of the test. Measures of internal consistency like Cronbach's α (Cronbach, 1951; Guttman, 1945) and McDonald's ω (McDonald, 1970) distill the reliability of a total score from the correlations among the items.

The equation for the reliability coefficient in Figure 5 shows a related but lesser known statistic called the *reliability index*, which is the correlation of an observed score with its true score. Squaring the reliability index yields the reliability coefficient. In regression analysis, the squared correlation between two variables is the proportion of variance one variable "explains" in the other. Thus, the reliability coefficient can be thought of as the proportion of variance in the observed score that is explained by the true score.

Figure 5

Reliability Is the Ratio of True Score Variance to Observed Score Variance



SEM ($\sigma_{X\,-\,T}$): The Typical Size of a Measurement Error

Because not everyone has the same score on a test, we need a statistic that describes how much people differ on a test, on average. Within a particular population, we can measure the population mean μ, and we can measure each *deviation*—how far each score X is from the mean μ by subtracting $X - \mu$. The typical size of the deviations is measured with a specialized kind of average called the *standard deviation*, symbolized with the Greek letter σ (see Appendix B and Figure B1 for technical details).

The SEM ($\sigma_{X\,-\,T}$) is the standard deviation of all measurement errors—the hypothetical fluctuations of observed scores X from a person's true score T. It is frequently symbolized as σ_e because $e = X - T$. The SEM can be thought of as the typical distance of observed scores from their respective true scores. In classical test theory, the SEM is assumed to be the same for everyone. In other measurement models, it differs depending on the observed score and possibly other characteristics of the person or situation (Raykov & Marcoulides, 2011).

If we knew everyone's observed and true scores, we could calculate the SEM directly by taking the standard deviation of the difference between true scores and the observed scores. However, if true scores were known, we would have no need to bother with observed scores and measurement error. Unfortunately, true scores are not available for inspection with observational data, and the direct calculation of the SEM is not possible. Nevertheless, we calculate the SEM indirectly via the reliability coefficient ρ_{XX}. See Appendix C for the derivation of the equation in Figure 6.

The SEM is the variation around a person's true score, as illustrated in Figure 7. The population mean for an IQ test is $\mu = 100$, with a population standard deviation of $\sigma = 15$. If a person's true score on the specific IQ test is $T = 70$, and the reliability of the score at IQ = 70 is $\rho_{XX} = .97$, the dark blue region represents the distribution of observed scores the person would obtain if the test could be readministered infinitely without carryover effects. The mean of this distribution is the true score $T = 70$, and the standard deviation (i.e., the SEM) is $\sigma_{X\,-\,T} \approx 2.60$, which, as seen in Equation 1, can be calculated by applying the equation in Figure 6:

$$\begin{aligned} \sigma_{X-T} &= \sigma_X\sqrt{1-\rho_{XX}} \\ 2.60 &\approx 15\sqrt{1-.97} \end{aligned} \quad . \tag{1}$$

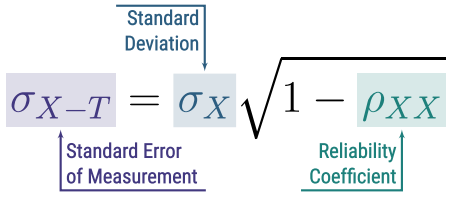
Confidence Intervals

With real observations, the true score is unknown, and its location is estimated from the observed score. After an IQ is calculated, it is usually presented along with the score's 95% confidence interval,

[3] In an extension of classical test theory called generalizability theory (Cronbach et al., 1963), the *universe score* is the average of all possible measurements of a particular construct, which means that it functions like a construct score but is still tied to actual measurements and will thus average in any flaws in measurement. Ideally, these flaws would average each other out and the universe score approaches the construct score. In this context, a *universe* refers to all possible conditions under which a measurement of a specific construct might occur, analogous to how *population* refers to all possible research subjects/persons in a specific group (Cronbach, 1972, p. 9).

Figure 6

The Standard Error of Measurement Is the Standard Deviation of the Difference Between Observed Scores X and True Scores T



the range that contains the true score in 95% of cases. If a person's observed IQ is 70, what is the best estimate of the range in which the person's true score is likely to be? If we know the reliability coefficient for a test at IQ = 70, we can use a regression equation to estimate the most likely range in which the true score is located.

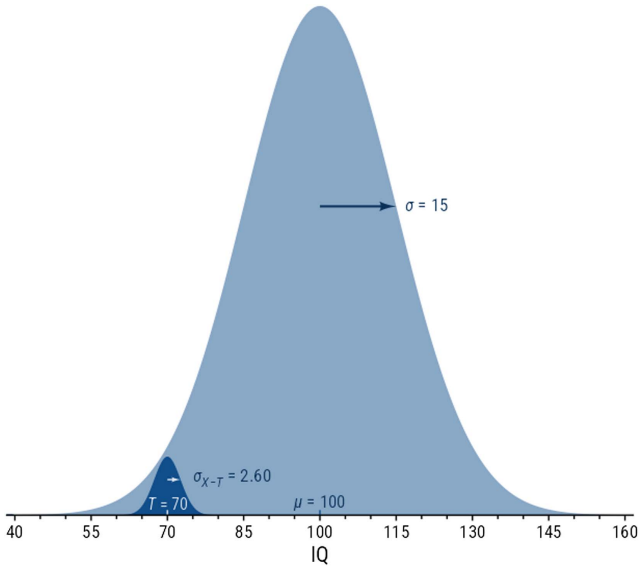
There are different kinds of confidence intervals that can answer different kinds of questions (Charter & Feldt, 2001; Levy & Mislevy, 2016; Stanley & Spence, 2024). Although court rulings often discuss confidence intervals based on the SEM (e.g., *Hall v. Florida*, 2014), most current tests use confidence intervals based on a closely related statistic called the *standard error of estimation*. Both types of confidence intervals answer valid questions and are correct equally often when their underlying assumptions hold.

If we somehow know the true score in advance, the SEM tells us the typical distance of the true score to the observed score. If we know the observed score—a much more likely event than knowing the true score—the standard error of estimation tells the typical distance from the true score to our best estimate of the true score.

The confidence interval based on the standard error of estimate is generally preferred for several reasons. First, it answers the question that clinicians and courts generally want to know: "For people with this observed score, where is their true score most likely to be?" By contrast, observed score–centered confidence intervals based on the SEM answer a question most of us do not have: "How wide does an interval centered on the observed need to be so that it will contain the true score 95% of the time we measure something?" The difference between those questions is subtle and abstract, but there is a plain and practical reason to prefer the confidence interval based on the standard error of estimation: Its interval is narrower. All else equal, a narrower prediction interval is more useful. Not only is the confidence interval narrower, but it is also usefully lopsided. When the observed score is very high or very low, the likely location of the true score is more likely to be closer to the population mean than it is to be even more extreme than the observed score. The confidence interval based on the SEM is symmetrically centered on the observed score, even when the location of the true score is not equally likely to be on either side.[4] Given a Bayesian interpretation, a confidence interval based on the standard error of estimation is also the 95% credible interval for the person's true score.

Figure 7

Variation of Observed Scores Around the True Score When T = 70, and the Reliability Coefficient Is .97



Note. The light blue region is the distribution of IQ test scores in the general population where the population mean is μ = 100 and the population standard deviation is σ = 15. The dark blue region is the distribution of IQ test scores that a person with a true score of $T = 70$ would obtain if the reliability of the scores is $\rho_{XX} = .97$. The mean of the dark blue region is T = true score with a standard error of measurement of $\sigma_{X-T} = 2.60$.

ID Criteria

The diverse forms of ID have many causes, some of which are gene and chromosomal variations that people either have or do not have (e.g., an additional copy of the 21st chromosome, which causes Down syndrome). Although some of the causes of ID involve categorical variables, the manifestation of difficulty is not categorical (i.e., present/absent) but comes in degrees of difficulty, which is why the diagnosis is accompanied by a severity specifier (mild, moderate, severe, or profound).

In the *DSM-5-TR*, there are three diagnostic criteria that must be met for a person to be given a diagnosis of ID:

1. Deficits in intellectual functions.
2. Deficits in adaptive functioning.

[4] An additional reason to prefer confidence intervals based on the standard error of estimation is that when reliability is low, confidence intervals based on the SEM are increasingly uninformative about the location of the true score, even though the true scores become increasingly easy to locate. The extreme case when the reliability coefficient is exactly zero illustrates this point well. When the reliability coefficient is zero, the equation in Figure 5 implies that the true score has no variance, and Equation C5 implies that the true score is precisely at the population mean. In this case, the equation in Figure 6 implies that the SEM is equal to the population standard deviation. The 95% confidence interval based on the SEM will still contain the true score 95% of the time, but the interval is so wide that it is no help in locating the true score, even though it is in a known, fixed position. By contrast, when the reliability coefficient is zero, Equation C2 implies that the confidence interval based on the standard error of estimation will also have a width of 0. Given a reliability of zero, this might be counterintuitive, but it is fitting because the true score has no variance, and its location is known to be exactly equal to the population mean. Thus, though both types of confidence intervals are correct as often as advertised, in the extreme case of zero reliability, the confidence interval based on the SEM is uninformative and the confidence interval based on the standard error of estimation hits the bulls' eye every time.

3. The onset of intellectual and adaptive deficits is observed during the development period (i.e., childhood and adolescence).

Based on accumulated clinical evidence, experience, and judgment, sensible cutoffs have been set for diagnostic thresholds, often with built-in buffer zones for cases in which flexibility is warranted. In the *DSM-5-TR*, deficits in intellectual functions must be confirmed by clinical assessment and by individualized, standardized intelligence testing. Because the 95% confidence interval for most comprehensive intelligence tests is about 10 points wide, the diagnostic threshold is roughly 70 ± 5.

The *DSM-5-TR* requires that intelligence test scores must not be the sole factor in determining deficits in intellectual functioning. The deficits must be evident in a comprehensive clinical assessment. Furthermore, the *DSM-5-TR* allows for a diagnosis of ID when scores are somewhat above the threshold when substantial adaptive behavior problems make the person's actual everyday functioning clinically comparable to individuals with IQ at or below the diagnostic range. In practice, individuals with observed IQ as high as 75 can be diagnosed with intellectual disability if their adaptive functioning deficits render their overall intellectual functioning to be comparable to individuals with observed IQ of 70 or below.

When Scores Fall on Both Sides of a Diagnostic Threshold

When an IQ is well below 70 or well above it, a second IQ test score will usually also fall on the same side of the diagnostic threshold. However, when individuals with a true score near the diagnostic threshold are given multiple tests, there is a substantial probability that at least some scores will fall on opposite sides of the threshold.

When faced with scores that are not on the same side of a diagnostic threshold, we can imagine a variety of methods to decide whether the diagnostic criteria have been met. Taking the mean value as the best estimate has intuitive appeal because the mean is normally an excellent measure of central tendency. Likewise, taking the median value might normally be a reasonable alternative if among the scores there are strong outliers of dubious accuracy.

Unfortunately, both the mean and the median introduce a known bias that is routinely corrected in most other contexts and ought to be corrected in this one as well.

Composite Score Calculations

To compute a composite score and its confidence interval, one needs the observed scores, population means, population standard deviations, and reliability coefficients of each test. The full correlation matrix between all scores is also needed. Although this article walks through the calculations one step at a time, all the calculations presented here are automated in a freely available spreadsheet at https://github.com/wjschne/assessingpsyche_resources/raw/main/CompositeIQFlynn.xlsx.

Consider the IQ results in Table 1. Almost all IQ tests have a mean of 100 and a standard deviation of 15, but there are exceptions. For example, the Stanford–Binet IV had a mean of 100 and a standard deviation of 16. Thus, Table 1 converts the Stanford–Binet IV IQ from its original metric to the common metric with a standard deviation of 15.

The unadjusted IQs are the typical scores that would normally be reported. The adjusted IQs are adjusted by the number of years elapsed from the norming date and the testing date multiplied by the size of the Flynn effect (Schalock et al., 2021; Floyd et al., 2021; McGrew, 2015; Watson, 2015). The Stanford–Binet IV and WISC-R are estimated to have a Flynn effect size of 2.94 points per decade (Trahan et al., 2014), and the WAIS-IV and WAIS-5 are estimated to have a Flynn effect of 1.2 points per decade (Wechsler et al., 2024; Winter et al., 2024).

Estimating Correlation Coefficients and Reliability Coefficients

To compute a composite score, the correlations among the scores need to be known within a plausible range. To compute the composite score's confidence interval, the reliability coefficient of each IQ must also be known.

Age-adjusted reliability coefficients are easily obtained from IQ test manuals, though it can sometimes be difficult to track down the reliability coefficients of lesser known tests now decades out of print. Often contemporary test reviews (e.g., from the Buros Center for Testing) contain such information. When a direct estimate is unavailable, internal consistency reliability coefficients for comprehensive IQ tests can safely be assumed to be in the range of .94 to .98, with .96 as a conservatively low point estimate for tests intended for adults and adolescents and .95 for tests intended for children.

Estimating correlations among test scores is not as straightforward as estimating reliability coefficients. Correlations between IQ measures are influenced by a variety of moderating variables, including the age of the person, the length of the retest interval, and the content emphases or domain sampling of the specific batteries in question (Breit et al., 2024). For this reason, direct estimates of correlations are preferred whenever possible. For example, the *WAIS-5 Technical Manual* (Wechsler et al., 2024) reports that from

Table 1
IQ Test Scores for a Hypothetical Individual Born on April 5, 1982

		Date		IQ			
Test	Age	Normed	Tested	Unadjusted	Adjusted	Flynn	r_{XX}
SB-IV[a]	6.5	1985	October 2, 1988	75.6	74.5	2.94	.96
WISC-R	8.0	1972	April 14, 1990	74.0	68.6	2.94	.96
WAIS-IV	36.3	2008	July 7, 2018	71.0	69.7	1.20	.98
WAIS-5	43.1	2024	April 30, 2025	75.0	74.8	1.20	.98

Note. SB-IV = Stanford–Binet Intelligence Scale, Fourth Edition; WISC-R = Wechsler Intelligence Scale, Revised; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition; WAIS-5 = Wechsler Adult Intelligence Scale, Fifth Edition.
[a] SB-IV IQ was 74 in its original metric with standard deviation = 16.

a sample of 186 individuals aged 16–90, the correlation of the Full Scale IQ from the WAIS-IV and the WAIS-5 was .92. The mean interval between administrations was 28 days (range = 7–134 days). If, for some reason, a person was given both the WAIS-IV and the WAIS-5 within a comparably short interval, then the .92 stability coefficient is a good guess as to what the correlation between the scores should be.

It is not always possible to find a large longitudinal study in which there are direct estimates of the stability coefficient for people with the same age, retest interval, and IQ battery pairing as needed to calculate the composite IQ for an individual. In this case, an estimate based on the best available evidence is necessary. For example, the meta-analytic models from Breit et al. (2024) provide reasonable estimates of IQ correlations based on data from 205 longitudinal studies of 87,408 participants total. Figure 8 shows the formula for estimating the stability coefficient for two 2 IQs, using the best fitting model for IQ from Breit et al.'s Table 3, adjusting for whether the longitudinal IQ pairs came from different batteries (e.g., WISC-III and Woodcock–Johnson III) or were the same test/same family (e.g., WISC-IV and WAIS-IV).

As shown in Figure 9, the stability coefficient is higher for adults than for children and adolescents. As expected, the stability coefficient decreases with longer retest intervals, though the decreases become smaller over time, and the stability coefficient curve approaches an asymptote. Each curve in Figure 9 is estimated to be lower by .104 if the two tests are from different test battery families (e.g., Wechsler and Stanford–Binet).

Assemble the Full Correlation Matrix of the Scores

The full correlation matrix of the scores includes the upper and lower halves of the matrix, as well as the ones on the diagonal. For the scores in Table 1, we can estimate the correlations using the equation in Figure 9. The resulting estimated stability coefficients (r) are shown in Table 2.

The correlations from Table 2 can be assembled into a full 4×4 correlation matrix R in Equation 2.

$$R = \begin{bmatrix} 1 & .67 & .52 & .52 \\ .67 & 1 & .65 & .65 \\ .52 & .65 & 1 & .79 \\ .52 & .65 & .79 & 1 \end{bmatrix}. \quad (2)$$

Choose a Weighting Scheme

Composite scores generally should be unit-weighted (i.e., all scores are equally weighted), as these are generally more robust than composites made with other weighting schemes (Wainer, 1976). Brief IQ batteries are often given for screening purposes or for periodic reevaluations after a comprehensive test has already been given. In isolation, brief IQ batteries are insufficiently reliable and comprehensive for diagnosing ID. If the brief score is already present in the person's records, it may be reasonable to include it in a composite IQ estimate, but it most likely should be given less weight than comprehensive IQs.

In rare situations, it is reasonable to downweight a test score if the score's accuracy is in question, but there is not enough evidence to throw the test score out completely. For example, the examiner's evaluation report might record that rapport was never fully established, and the examiner could not tell if the examinee's best effort was obtained. To keep things simple and consistent, we recommend that any test that is downweighted be given a weight of .5 instead of 1 if there remains justification for retaining the score at all.

The weights can be placed in a vector (i.e., sequence of numbers) called w. In the most common application, vector w consists entirely of ones as in Equation 3.

$$\begin{aligned} w &= \{w_1, w_2, w_3, w_4\} \\ &= \{1, 1, 1, 1\}. \end{aligned} \quad (3)$$

If Test 3 is given a weight of .5, the weight vector will look like Equation 4.

$$w = \{1, 1, .5, 1\}. \quad (4)$$

A Weighted Sum

A simple sum is simple enough. A weighted sum gives more weight to some elements than others. In matrix notation, if X is a vector of k scores and w is a vector of k weights, then a weighted sum is $w'X$ (see Equation 5).

Figure 8

Predicted Stability Coefficient as a Function of Age, Retest Interval, and Battery Difference

$$r = .716 - .003e^{-.258(A-20)} + .095e^{-.138(I-5)} - .104D$$

Stability Coefficient; Age (in years); Retest Interval (in years); Different Battery Family (1 = Different; 0 = Same)

Figure 9

Stability Coefficients Based on Best Fitting Model for General Ability From Table 3 in Breit et al. (2024)

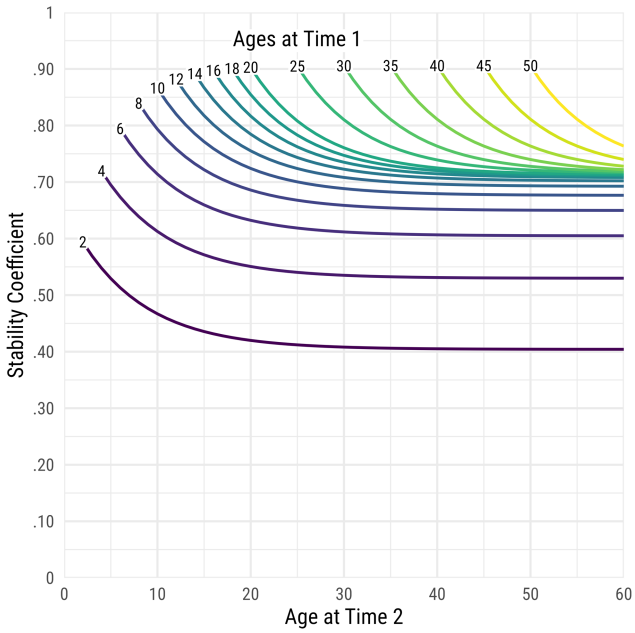


Table 2
Estimated Stability Coefficients (r) for Hypothetical Individual

Test			Age			
Time 1	Time 2	Family	Time 1	Time 2	Interval	Estimated *r*
SB-IV	WISC-R	Different	6.5	8.0	1.5	.67
SB-IV	WAIS-IV	Different	6.5	36.3	29.8	.52
SB-IV	WAIS-5	Different	6.5	43.1	36.6	.52
WISC-R	WAIS-IV	Same	8.0	36.3	28.2	.65
WISC-R	WAIS-5	Same	8.0	43.1	35.0	.65
WAIS-IV	WAIS-5	Same	36.3	43.1	6.8	.79

Note. SB-IV = Stanford–Binet Intelligence Scale, Fourth Edition; WISC-R = Wechsler Intelligence Scale, Revised; WAIS-IV = Wechsler Adult Intelligence Scale, Fourth Edition; WAIS-5 = Wechsler Adult Intelligence Scale, Fifth Edition.

$$
\begin{aligned}
w'X &= \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_k \end{bmatrix}' \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \\
&= \begin{bmatrix} w_1\, w_2 \cdots w_k \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \\
&= w_1X_1 + w_2X_2 + \ldots + w_kX_k.
\end{aligned} \tag{5}
$$

In the computation of a weighted composite score, we need to compute the weighted sum of score deviations from the population mean μ_X (see Equation 6).

$$
\begin{aligned}
w'(X - \mu) = {} & w_1(X_1 - \mu_1) + w_1(X_2 - \mu_2) + \ldots \\
& + w_k(X_k - \mu_k).
\end{aligned} \tag{6}
$$

In our hypothetical example, the weights are all ones. Thus, for the unadjusted IQs, the sum of the deviations is computed as in Equation 7.

$$
\begin{aligned}
w'(X - \mu) = {} & 1(75.6 - 100) + 1(74 - 100) + 1(71 - 100) \\
& + 1(75 - 100) \\
= {} & -104.4.
\end{aligned} \tag{7}
$$

For the adjusted IQs, the sum of the deviations is −112.3.

The Standard Deviation of a Weighted Sum

The standard deviation of the sum of many variables is the square root of the sum of the variables' covariance matrix. The covariance between a single variable X and a single variable Y is $\sigma_{XY} = \sigma_X\sigma_Y r_{XY}$. For a vector of many variables $X = \{X_1 + X_2 + \ldots + X_k\}$ with standard deviations in vector $\sigma = \{\sigma_1 + \sigma_2 + \ldots + \sigma_k\}$, the covariance matrix Σ is comptuted as is Equation 8.

$$
\Sigma = \text{diag}(\sigma)R\ \text{diag}(\sigma). \tag{8}
$$

The diag operator converts a vector into a diagonal matrix (i.e., the vector is in the diagonal of the matrix, and all other values are 0). In our hypothetical example, the covariance matrix is computed as in Equation 9:

$$
\begin{aligned}
\Sigma &= \text{diag}(\sigma)R\ \text{diag}(\sigma) \\
&= \begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & 15 \end{bmatrix} \begin{bmatrix} 1 & .67 & .52 & .52 \\ .67 & 1 & .65 & .65 \\ .52 & .65 & 1 & .79 \\ .52 & .65 & .79 & 1 \end{bmatrix} \begin{bmatrix} 15 & 0 & 0 & 0 \\ 0 & 15 & 0 & 0 \\ 0 & 0 & 15 & 0 \\ 0 & 0 & 0 & 15 \end{bmatrix} \\
&= \begin{bmatrix} 225.00 & 150.18 & 116.39 & 115.96 \\ 150.18 & 225.00 & 147.14 & 146.61 \\ 116.39 & 147.14 & 225.00 & 177.73 \\ 115.96 & 146.61 & 177.73 & 225.00 \end{bmatrix}.
\end{aligned} \tag{9}
$$

The standard deviation of a weighted sum is the sum of the weighted covariance matrix, calculated as in Equation 10.

$$
\sigma_S = \sqrt{w'\Sigma w}. \tag{10}
$$

The expression $w'\Sigma w$ means "make a weighted sum of the matrix Σ." In a unit-weighted composite, the expression reduces to the simple sum of the entire covariance matrix. In the current example, $\sigma_s \approx 51.07$.

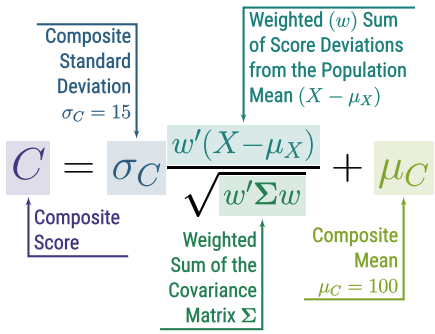
Computing the Composite IQ

The equation in Figure 10 for computing a composite score might look complicated at first, but it is not so difficult after spending time with it. If the composite is unit-weighted and all scores have a mean of 100 and a standard deviation of 15, the formula becomes even easier—sum the score deviations from 100, divide by the square root of the sum of the correlation matrix, and add 100.

In our hypothetical example, the weighted sum of the unadjusted IQ deviations was calculated using Equation 6 to be −104.4, and the square root of the weighted sum of the covariance matrix was calculated using Equation 10 to be 51.07. Thus, the unadjusted composite score is computed with Equation 11.

$$
\begin{aligned}
C &= \sigma_C \frac{w'(X - \mu_X)}{\sqrt{w'\Sigma w}} + \mu_C \\
&\approx 15\frac{-104.4}{51.07} + 100 \\
&\approx 69.3.
\end{aligned} \tag{11}
$$

Figure 10
Formula for Weighted Composite Scores



Applying the formula from Figure 10, the composite IQ using adjusted IQs is 67.0.

Computing the Weighted Composite Confidence Interval

To compute the confidence interval, the composite reliability coefficient is needed first. The equation for the reliability coefficient of the composite in Figure 11 is simpler than it looks. The denominator is the weighted sum of the correlation matrix. The numerator is the same as the denominator except that the diagonal of ones has been replaced with the vector of the tests' reliability coefficients.

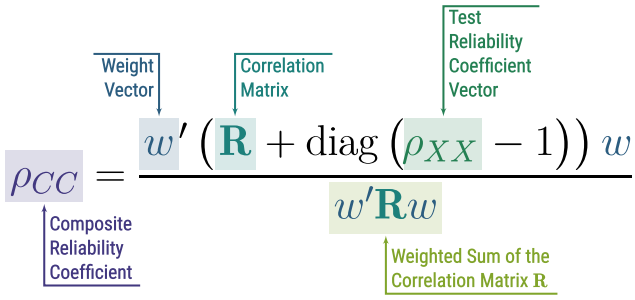
In our example, the reliability coefficient is computed with Equation 12.

$$\begin{aligned}\rho_{CC} &= \frac{w'(R + \operatorname{diag}(\rho_{XX} - 1))w}{w'Rw} \\ &= \frac{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}' \begin{bmatrix}.96 & .67 & .52 & .52\\ .67 & .96 & .65 & .65\\ .52 & .65 & .98 & .79\\ .52 & .65 & .79 & .98\end{bmatrix} \begin{bmatrix}1\\1\\1\\1\end{bmatrix}}{\begin{bmatrix}1\\1\\1\\1\end{bmatrix}' \begin{bmatrix}1 & .67 & .52 & .52\\ .67 & 1 & .65 & .65\\ .52 & .65 & 1 & .79\\ .52 & .65 & .79 & 1\end{bmatrix} \begin{bmatrix}1\\1\\1\\1\end{bmatrix}} \\ &\approx .9896.\end{aligned} \tag{12}$$

The standard error of estimation of the composite C is obtained with Equation C2, and the 95% confidence interval is constructed by applying the equation in Figure C3 (see Appendix C) as seen in Equation 13.

$$\begin{aligned}95\%\ \text{CI} &= \rho_{CC}(C - \mu_C) + \mu_C \pm z_{95}\sigma_C\sqrt{\rho_{CC} - \rho_{CC}^2} \\ &\approx .9896(69.3 - 100) + 100 \pm 1.96 \cdot 15\sqrt{.9896 - .9896^2} \\ &\approx 69.66 \pm 2.98 \\ &\approx [67,73].\end{aligned} \tag{13}$$

Figure 11
Formula for the Reliability Coefficient of a Weighted Composite



The 95% confidence interval for the composite IQ adjusted for norm obsolescence is approximately 64–70.

Because composite scores are generally more reliable than the individual scores they summarize, their confidence intervals are narrower as well. As seen in Figure 12, the confidence interval width narrows as the number of scores increases. The width of confidence interval also depends on how reliable the individual scores are on average and the average correlation among the scores. For example, a single score with a reliability of .97 has a 95% confidence interval that is 10 points wide. If five such scores with an average intercorrelation of .8 are combined into a composite score, the confidence interval width narrows to half its original size.

Recommendations for Creating Composite IQs

The process of computing an overall composite IQ requires much more than the simple application of formulas. Before any computations are possible, a wide range of nonquantitative influences on IQ test results must be considered and interpreted via the careful, consistent, and rigorous application of clinical judgment and professional ethics. We present here a set of recommendations about the nonquantitative decisions and interpretations that precede computation of overall composite IQs.

Consider the Influence of Carryover Effects

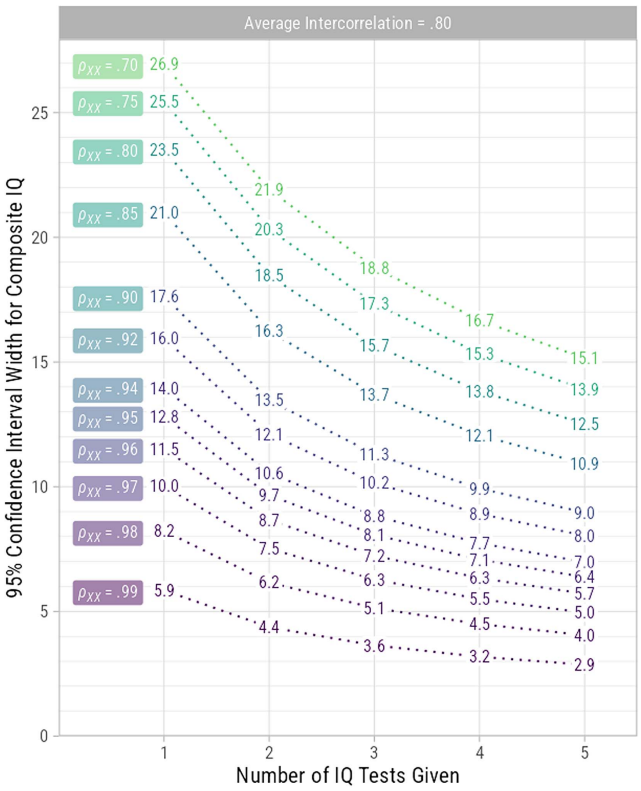
Carryover effects describe how past assessments influence subsequent assessments negatively or positively. Carryover effects can occur within testing sessions and between testing sessions.

Fatigue effects, broadly conceived, refer to situations in which previous testing reduces a person's ability and/or motivation to perform. Experienced clinicians set the conditions for optimal test performance and closely monitor examinees for signs of fatigue, discouragement, and loss of motivation to perform. Most evaluation reports include statements by the clinician about whether the examinee was willing and able to give the test their best effort. If the clinician does not believe that the examinee performed to the best of their ability, the test should probably not be combined with a score obtained under more favorable circumstances. An emerging standard of practice in clinical contexts is now to include measures of performance validity as a component of evaluations that include maximum performance ability tests (versus typical performance measures like adaptive behavior; Cronbach, 1994) such as measures of IQ, achievement, memory, or other neuropsychological functions (Sweet et al., 2021).

Practice effects, broadly conceived, refer to ways in which having been tested before results in higher scores than would otherwise have been obtained. According to the *APA Dictionary of Psychology* (VandenBos, 2007), a practice effect is "Any change or improvement that results from practice or repetition of task items or activities"

Figure 12

The Effect of Single-Test Reliability and the Number of Tests in the Composite on the Composite's Confidence Interval Width



(p. 719). For example, tests that give time bonuses to solve puzzle-like tasks can be solved more quickly the second time if one simply remembers the solution (or a previously used task-specific strategy) rather than working it out anew. Tests with manipulable parts typically have larger practice effects than nonmanipulable tasks. Practice effects are not to be misunderstood as an individual engaging in actual deliberate "practice" on intelligence test items or tasks. Instead, it refers to increases to scores from the repeated exposure to test items and tasks that are not due to a true change in a person's abilities (Heilbronner et al., 2010). The explanation is that some form of incidental learning occurs during the taking of an IQ test, most typically involving performance or nonverbal subtests (Greenspan & Olley, 2015; Kaufman & Lichtenberger, 2006). The practice effect professional and research literature is large (Goldstein & Saklofske, 2010; Kaufman & Lichtenberger, 2006; Scharfen et al., 2018) and cannot be covered in depth in this article.

In general practice, effects on IQ tests over short intervals (3–6 months) are about 4–7 points, on average (Basso et al., 2002; Estevis et al., 2012; Wechsler, 2008; Wechsler et al., 2024). Longer term intervals (1–3 years) are associated with minimal practice effects in children, particularly with verbal measures (Watkins & Smith, 2013; Watkins et al., 2022), and 1–3 points in adults (Bartels et al., 2010). After 5–7 years, most practice effects are minimal in adults (Calamia et al., 2012). Furthermore, practice effects tend to be smaller and shorter lived in older adults and in clinical groups that tend to score low on IQ tests (Calamia et al., 2012; Edgin et al., 2017; Jutten et al., 2020; Winter et al., 2024).

Currently, there are no scientific or professionally accepted methods or equations for adjusting scores for practice effects. Instead, if it is plausible that practice effects should alter the interpretation of the scores, examiners should present the original scores and an estimate of how much the practice effect might have influenced the scores. For example, for an individual given the same test twice, 1 month apart:

> Mr. Johnson's Full Scale IQ on the second administration of the WAIS-5 was 76, 6 points higher than the Full Scale IQ of 70 on the first administration. When the same intelligence test is given twice over a short time period, most people score higher on the second administration, often because they recall the test items and can perform the tasks more quickly. The various reasons people tend to score higher on a second round of testing are known collectively as *practice effects*. Over short intervals of a few months, practice effects often result in IQ results increasing by about 4–7 points, on average. Thus, Mr. Johnson's second Full Scale IQ result is well within expectations and should be regarded as essentially equivalent to the first result.

Consider the Influence of Norm Obsolescence

Norms for ability tests do not necessarily stay fixed across generations. For example, U.S. soldiers in World War II performed much better than did soldiers in World War I on the Army General Classification Test (Tuddenham, 1948). Indeed, for most of the 20th century, IQ test scores increased steadily in the United States and over 70 countries in which the phenomenon was studied (Flynn, 1984, 1987; Trahan et al., 2014; Wongupparaj et al., 2023). Generational increases in cognitive ability scores are called "Flynn effects" in honor of James Flynn, who first presented systematic and persuasive evidence of this widespread, long-standing trend. Psychologists and psychological measurement experts typically describe the Flynn effect as resulting from a "softening" of IQ test norms with the passage of time that results in a person's IQ test performance "comparison to a historical reference group from the past—not the person's contemporary peers" (McGrew, 2015, p. 155).

The size of Flynn effects varies over time, place, population, and ability type. Although cognitive scores, on average, continue to increase steadily in developing and middle-income countries, there is increasing evidence that the growth in scores may have slowed in recent decades in the United States and may have reversed in parts of northern Europe (Dutton et al., 2016; Dworak et al., 2023; Platt et al., 2019; Wongupparaj et al., 2023). Whereas growth in the Wechsler intelligence tests held steady at about 3 points per decade until the WAIS-IV and Wechsler Intelligence Scale for Children, Fifth Edition (Weiss et al., 2015), the growth from the WAIS-IV to the WAIS-5 was only 1.2 points per decade (Winter et al., 2024).

Because ability test norms tend to shift over the decades, they are revised and renormed regularly to minimize distortions in diagnosis and classification rates (Kanaya & Ceci, 2010). Before it was clear how consistent Flynn effects were in the 20th century, intelligence tests were revised less frequently than is typical today. For example, 26 years elapsed between the first edition of the WAIS (1955) and the Wechsler Adult Intelligence Scale, Revised (1981). If scores rose 3 points per decade, a person given the 1955 WAIS in 1981 in the last year of its lifecycle would be predicted to score 7.8 points higher than if the person had been given the then-new Wechsler Adult Intelligence Scale, Revised. To minimize such interpretive difficulties, most publishers of IQ tests now make efforts to revise tests or provide updated norms within 10–15 years of publication.

When making lower stakes decisions, a clinician can informally manage distortion due to norm obsolescence (McGrew, 2015). In capital cases, every point counts, and the effects of norm obsolescence must be considered in a rigorous, systematic way. In most contexts, psychologists do not present scores corrected for the Flynn effect (Hagan et al., 2008). Usually, there is no need for corrected scores, and typically, an accurate estimate of the size of the correction is not known until the test is already outdated. A capital case is one of the few contexts in which it is both vital and possible to calculate accurate score corrections (Schalock et al., 2021; Floyd et al., 2021; McGrew, 2015; Watson, 2015).

Forensic psychologists working on capital cases routinely evaluate scores from superannuated tests in which the size of the Flynn effect for that test is known. Grégoire et al. (2016) recommended using direct estimates of Flynn effect sizes whenever possible. For example, the WAIS-5 Technical and Interpretive Manual (Wechsler et al., 2024) reports that the size of the Flynn effect in the change from the WAIS-IV to the WAIS-5 is 0.012 points per year. Grégoire et al. also noted that the size of the Flynn effect can depend on a large number of factors (e.g., age and ability level) that should be taken into account when appropriate, in accordance with the best available evidence. Even though it may be difficult to estimate for some tests and over time, as the size of the Flynn effect may change, application of the Flynn correction in high-stakes decision making should be implemented to obtain the most accurate estimates of intellectual function possible (Floyd et al., 2021; McGrew, 2015; Reynolds et al., 2010; Watson, 2015).

When direct estimates of Flynn effect sizes are not available, up-to-date meta-analytic estimates are reasonable substitutes (Grégoire et al., 2016). For IQ from comprehensive intelligence tests up to the first decade of the current century in the United States, the Flynn effect was about .294 points per year (Trahan et al., 2014). In the past 2 decades, it is likely that the rate of growth has slowed (Winter et al., 2024) and may be in the process of reversing (Dworak et al., 2023). Even so, in some samples, it appears that meaningful gains are still occurring at the lower end of the ability distribution (Flynn & Shayer, 2018; Oberleiter et al., 2024). Anyone practicing the diagnosis of ID needs to stay contemporary with the continuously developing Flynn effect literature and apply the best available research outcomes in practice.

Combine Scores That Are Conceptually Coherent

A composite score needs to consist of conceptually related scores intended to measure the same attribute or closely related functions. Individually administered, nationally representative, comprehensive general intelligence tests are measures of global cognitive functioning and should only be combined with other similar measures of global cognitive functioning. Intelligence consists of at least eight interrelated broad abilities, possibly more (Carroll, 1993; Schneider & McGrew, 2018). Comprehensive tests not only need to be sufficiently reliable to permit high-stakes decisions about individuals, but they also need to include a diverse set of measures of multiple broad abilities. It is recommended that comprehensive tests of general intelligence sample at least three (preferably more) broad abilities (Schalock et al., 2021; Floyd et al., 2021).

Clinicians often administer standalone measures of specific cognitive abilities within the domains of attention, memory, perception, executive functions, language, and knowledge to understand the nature of a person's strengths and difficulties. Many people, including individuals with ID, have one or more specific ability scores well above or well below their IQ (Bergeron & Floyd, 2006, 2013; Bergeron et al., 2023). By definition, specific ability scores are not designed to measure overall intellectual capacity. Thus, although intelligence tests include measures of specific cognitive abilities, IQ should not be combined into a single composite score with standalone measures of specific cognitive functions (e.g., a global IQ combined with a measure of visual working memory).

Only Combine Scores Deemed to Be Validly Obtained

IQ results can be compromised for a variety of reasons, including malingering, intoxication, fatigue, brief psychosis, administration errors, and poor rapport with the examiner. Scores believed to be inaccurate ought not to be combined into a composite score.

Combine Scores Deemed to Be Comparable Estimates of Overall Intellectual Functioning

Intelligence test scores should be combined only when they are thought to estimate the person's long-standing overall capacity to reason, learn, and adapt. If we have reason to believe that a person's overall intelligence has changed dramatically because of a head injury, for example, it would be a mistake to combine scores obtained before the injury with scores obtained after the injury. Likewise, scores before and after the onset of dementia, schizophrenia, or other brain diseases known to lower cognitive capacity should not be combined.

Limitations and Future Directions

Ideally, future research will allow for more precise estimates of norm obsolescence, practice effects, and stability coefficients conditioned on relevant conditions.

Conclusion

The diagnosis of ID is a profound and life-changing event. The process of determining ID status must use the best available methods. When multiple IQ results are available, the information must be combined in a rigorous, systematic way. When so combined into a composite score, this score should be interpreted in the context of expert-based clinical judgment. We have demonstrated that when relevant and comparable IQ results are combined, they should be combined into proper composite scores, like how IQ itself or any other composite score is constructed. We have emphasized the application of this method of combining scores to the diagnosis of ID; however, the method is applicable to any set of scores where it would be informative to know what the composite of these scores would reveal and when it is reasonable to combine the scores from the set of tests being considered.

References

American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders (DSM-5-TR).* https://doi.org/10.1176/appi.books.9780890425787

Atkins v. Virginia, 536 U.S. 304 (2002). https://www.oyez.org/cases/2001/00-8452

Bartels, C., Wegrzyn, M., Wiedl, A., Ackermann, V., & Ehrenreich, H. (2010). Practice effects in healthy adults: A longitudinal study on frequent repetitive cognitive testing. *BMC Neuroscience*, *11*(1), Article 118. https://doi.org/10.1186/1471-2202-11-118

Basso, M. R., Carona, F. D., Lowery, N., & Axelrod, B. N. (2002). Practice effects on the WAIS-III Across 3- and 6-Month Intervals. *The Clinical Neuropsychologist*, *16*(1), 57–63. https://doi.org/10.1076/clin.16.1.57.8329

Bergeron, R., & Floyd, R. G. (2006). Broad cognitive abilities of children with mental retardation: An analysis of group and individual profiles. *American Journal on Mental Retardation*, *111*(6), 417–432. https://doi.org/10.1352/0895-8017(2006)111[417:BCAOCW]2.0.CO;2

Bergeron, R., & Floyd, R. G. (2013). Individual part score profiles of children with intellectual disability: A descriptive analysis across three intelligence tests. *School Psychology Review*, *42*(1), 22–38. https://doi.org/10.1080/02796015.2013.12087489

Bergeron, R., Floyd, R. G., McNicholas, P. J., & Farmer, R. L. (2023). Assessment of intellectual disability with the Wechsler Intelligence Scale for Children, fifth edition: Analysis of part score profiles and diagnostic outcomes. *School Psychology Review*, *52*(6), 747–762. https://doi.org/10.1080/2372966X.2022.2094284

Bienaymé, I.-J. (1853). Considérations à l'appui de la découverte de Laplace sur la loi de probabilité dans la méthode des moindres carrés [Considerations in support of Laplace's discovery concerning the probability law in the method of least squares]. *Comptes Rendus de l'Académie Des Sciences Paris*, *37*(309–317), 309–317.

Blume, J. H., Johnson, S. L., & Seeds, C. (2009). An empirical look at Atkins v. Virginia and its application in capital cases. *Tennessee Law Review*, *76*(3), 625–639. https://scholarship.law.cornell.edu/facpub/7

Borsboom, D., & Mellenbergh, G. J. (2002). True scores, latent variables, and constructs. *Intelligence*, *30*(6), 505–514. https://doi.org/10.1016/S0160-2896(02)00082-X

Botev, Z. I., Mandjes, M., & Ridder, A. (2015). Tail distribution of the maximum of correlated Gaussian random variables. *2015 Winter Simulation Conference (WSC)* (pp. 633–642). https://doi.org/10.1109/WSC.2015.7408202

Breit, M., Scherrer, V., Tucker-Drob, E. M., & Preckel, F. (2024). The stability of cognitive abilities: A meta-analytic review of longitudinal studies. *Psychological Bulletin*, *150*(4), 399–439. https://doi.org/10.1037/bul0000425

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Calamia, M., Markon, K., & Tranel, D. (2012). Scoring higher the second time around: Meta-analyses of practice effects in neuropsychological assessment. *The Clinical Neuropsychologist*, *26*(4), 543–570. https://doi.org/10.1080/13854046.2012.680913

Carlson, S. R., Morningstar, M. E., & Munandar, V. (2020). Workplace supports for employees with intellectual disability: A systematic review of the intervention literature. *Journal of Vocational Rehabilitation*, *52*(3), 251–265. https://doi.org/10.3233/JVR-201075

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge University Press. https://doi.org/10.1017/cbo9780511571312

Cavanagh, J., Meacham, H., Pariona-Cabrera, P., & Bartram, T. (2021). Subtle workplace discrimination inhibiting workers with intellectual disability from thriving at the workplace. *Personnel Review*, *50*(7/8), 1739–1756. https://doi.org/10.1108/PR-10-2021-0723

Charter, R. A., & Feldt, L. S. (2001). Confidence intervals for true scores: Is there a correct approach? *Journal of Psychoeducational Assessment*, *19*(4), 350–364. https://doi.org/10.1177/073428290101900404

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/bf02310555

Cronbach, L. J. (Ed.). (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

Cronbach, L. J. (1994). *Essentials of psychological testing* (5th ed.). HarperCollins Publishers.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Dennett, D. C. (2013). *Intuition pumps and other tools for thinking*. Allen Lane.

Dutton, E., Linden, D. V., & Lynn, R. (2016). The negative Flynn effect: A systematic literature review. *Intelligence*, *59*, 163–169. https://doi.org/10.1016/j.intell.2016.10.002

Dworak, E. M., Revelle, W., & Condon, D. M. (2023). Looking for Flynn effects in a recent online U.S. adult sample: Examining shifts within the SAPA project. *Intelligence*, *98*, Article 101734. https://doi.org/10.1016/j.intell.2023.101734

Edgin, J. O., Anand, P., Rosser, T., Pierpont, E. I., Figueroa, C., Hamilton, D., Huddleston, L., Mason, G., Spanò, G., Toole, L., Nguyen-Driver, M., Capone, G., Abbeduto, L., Maslen, C., Reeves, R. H., & Sherman, S. (2017). The Arizona Cognitive Test battery for down syndrome: Test–retest reliability and practice effects. *American Journal on Intellectual and Developmental Disabilities*, *122*(3), 215–234. https://doi.org/10.1352/1944-7558-122.3.215

Estevis, E., Basso, M. R., & Combs, D. (2012). Effects of practice on the Wechsler Adult Intelligence Scale-IV across 3- and 6-month intervals. *The Clinical Neuropsychologist*, *26*(2), 239–254. https://doi.org/10.1080/13854046.2012.659219

Floyd, R. G., Farmer, R. L., Schneider, W. J., & McGrew, K. S. (2021). Theories and measurement of intelligence. In L. Glidden, L. Abbeduto, L. L. McIntyre, & M. J. Tassé (Eds.), *APA handbook of intellectual and developmental disabilities* (Vol. 1, pp. 386–424). American Psychological Association. https://doi.org/10.1037/0000194-015

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*(1), 29–51. https://doi.org/10.1037/0033-2909.95.1.29

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191. https://doi.org/10.1037/0033-2909.101.2.171

Flynn, J. R., & Shayer, M. (2018). IQ decline and Piaget: Does the rot start at the top? *Intelligence*, *66*, 112–121. https://doi.org/10.1016/j.intell.2017.11.010

Furr, R. M. (2022). *Psychometrics: An introduction* (4th ed.). Sage Publications.

Goldstein, G., & Saklofske, D. H. (2010). The Wechsler Intelligence Scales in the assessment of psychopathology. In L. G. Weiss, D. H. Saklofske, D. L. Coalson, & S. E. Raiford (Eds.), *WAIS-IV clinical use and interpretation* (pp. 189–216). Elsevier. https://doi.org/10.1016/B978-0-12-375035-8.10007-2

Gormley, M. E. (2015). Workplace stigma toward employees with intellectual disability: A descriptive Study1. *Journal of Vocational Rehabilitation*, *43*(3), 249–258. https://doi.org/10.3233/JVR-150773

Greenspan, S., & Olley, J. G. (2015). Variability in IQ test scores. In E. A. Polloway (Ed.), *The death penalty and intellectual disability* (pp. 41–153). American Association on Intellectual and Developmental Disability.

Grégoire, J., Daniel, M., Llorente, A. M., & Weiss, L. G. (2016). The Flynn effect and its clinical implications. In L. G. Weiss, D. H. Saklofske, J. A. Holdnack, & A. Prifitera (Eds.), *WISC-V Assessment and Interpretation* (pp. 187–212). Elsevier. https://doi.org/10.1016/B978-0-12-404697-9.00006-6

Guttman, L. (1945). A basis for analyzing test–retest reliability. *Psychometrika*, *10*(4), 255–282. https://doi.org/10.1007/bf02288892

Hagan, L. D., Drogin, E. Y., & Guilmette, T. J. (2008). Adjusting IQ scores for the Flynn effect: Consistent with the standard of practice? *Professional Psychology: Research and Practice*, *39*(6), 619–625. https://doi.org/10.1037/a0012693

Hall v. Florida, 572 U.S. 701 (2014). https://supreme.justia.com/cases/federal/us/572/701/

Hamm v. Smith, 604 U.S. ___ (2024). https://supreme.justia.com/cases/federal/us/604/23-167/

Heilbronner, R. L., Sweet, J. J., Attix, D. K., Krull, K. R., Henry, G. K., & Hart, R. P. (2010). Official position of the American Academy of Clinical Neuropsychology on serial neuropsychological assessments: The Utility and Challenges of Repeat Test administrations in clinical and forensic contexts. *The Clinical Neuropsychologist*, *24*(8), 1267–1278. https://doi.org/10.1080/13854046.2010.526785

Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik [The actual content of quantum theoretical kinematics and mechanics]. *Zeitschrift für Physik*, *43*(3–4), 172–198. https://doi.org/10.1007/BF01397280

Horst, P. (1936). Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, *1*(1), 53–60. https://doi.org/10.1007/BF02287924

Jutten, R. J., Grandoit, E., Foldi, N. S., Sikkes, S. A. M., Jones, R. N., Choi, S., Lamar, M. L., Louden, D. K. N., Rich, J., Tommet, D., Crane, P. K., & Rabin, L. A. (2020). Lower practice effects as a marker of cognitive performance and dementia risk: A literature review. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, *12*(1), Article e12055. https://doi.org/10.1002/dad2.12055

Kanaya, T., & Ceci, S. (2010). The impact of the Flynn effect on LD diagnoses in special education. *Journal of Learning Disabilities*, *45*(4), 319–326. https://doi.org/10.1177/0022219410392044

Kaufman, A. S., & Lichtenberger, E. O. (2006). *Assessing adolescent and adult intelligence* (3rd ed.). Wiley.

Levy, R., & Mislevy, R. J. (2016). *Bayesian psychometric modeling*. Taylor & Francis Group. https://doi.org/10.1201/9781315374604

Ley, P. (1972). *Quantitative aspects of psychological assessment: An introduction*. Duckworth.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test scores*. Addison-Wesley.

Lysaght, R., Ouellette-Kuntz, H., & Lin, C.-J. (2012). Untapped potential: Perspectives on the employment of people with intellectual disability. *WORK: A Journal of Prevention, Assessment & Rehabilitation*, *41*(4), 409–422. https://doi.org/10.3233/WOR-2012-1318

Madan, S., Kumar, T., & Bhagat, A. (2024). Inclusive hiring through technology: A recruitment platform for individuals with intellectual disabilities. *Procedia CIRP*, *128*, 31–36. https://doi.org/10.1016/j.procir.2024.05.091

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McGrew, K. S. (2015). Norm obsolescence: The Flynn Effect. In E. A. Polloway (Ed.), *The death penalty and intellectual disability* (pp. 155–169). American Association on Intellectual and Developmental Disabilities. https://www.aaidd.org/publications/bookstore-home/product-listing/the-death-penalty-and-intellectual-disability

Oberleiter, S., Fries, J., Dejardin, F., Heller, J., Schaible, C., Vetter, M., Voracek, M., & Pietschnig, J. (2024). Inconsistent Flynn effect patterns may be due to a decreasing positive manifold: Cohort-based measurement-invariant IQ test score changes from 2005 to 2024. *Intelligence*, *107*, Article 101867. https://doi.org/10.1016/j.intell.2024.101867

Paunonen, S. V. (1984). Optimizing the validity of personality assessments: The importance of aggregation and item content. *Journal of Research in Personality*, *18*(4), 411–431. https://doi.org/10.1016/0092-6566(84)90001-1

Platt, J. M., Keyes, K. M., McLaughlin, K. A., & Kaufman, A. S. (2019). The Flynn effect for fluid IQ may not generalize to all ages or ability levels: A population-based study of 10,000 US adolescents. *Intelligence*, *77*, Article 101385. https://doi.org/10.1016/j.intell.2019.101385

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Reynolds, C. R., Altmann, R. A., & Allen, D. N. (2021). *Mastering modern psychological testing: Theory and methods*. Springer International Publishing. https://doi.org/10.1007/978-3-030-59455-8

Reynolds, C. R., & Kamphaus, R. W. (2026). *Reynolds Intellectual Assessment Scales–Second edition normative update*. Psychological Assessment Resources.

Reynolds, C. R., Niland, J., Wright, J. E., & Rosenn, M. (2010). Failure to apply the Flynn correction in death penalty litigation: Standard practice of today maybe, but certainly malpractice of tomorrow. *Journal of Psychoeducational Assessment*, *28*(5), 477–481. https://doi.org/10.1177/0734282910373348

Schalock, R. L., Luckasson, R., & Tassé, M. J. (2021). *Intellectual disability: Definition, diagnosis, classification, and systems of supports*. American Association on Intellectual and Developmental Disabilities.

Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. *Intelligence*, *67*, 44–66. https://doi.org/10.1016/j.intell.2018.01.003

Schneider, W. J. (2013). Principles of assessment of aptitude and achievement. In D. Saklofske, C. R. Reynolds, & V. Schwean (Eds.), *The Oxford handbook of child psychological assessment* (pp. 286–330). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199796304.013.0013

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 (7)*. Houghton Mifflin Harcourt. https://info.riversideinsights.com/hubfs/ASBs/WJIV_ASB_7_FINAL.pdf

Schneider, W. J., Flanagan, D. P., Niileksela, C. R., & Engler, J. R. (2024). The effect of measurement error on the positive predictive value of PSW methods for SLD identification: How buffer zones dispel the illusion of inaccuracy. *Journal of School Psychology*, *103*, Article 101280. https://doi.org/10.1016/j.jsp.2023.101280

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of intelligence. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (4th ed., pp. 73–130). Guilford Press. https://www.guilford.com/books/Contemporary-Intellectual-Assessment/Flanagan-McDonough/9781462552030

Spearman, C. E. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stanley, D. J., & Spence, J. R. (2024). The comedy of measurement errors: Standard error of measurement and standard error of estimation. *Advances in Methods and Practices in Psychological Science*, *7*(4). https://doi.org/10.1177/25152459241285885

Sweet, J. J., Heilbronner, R. L., Morgan, J. E., Larrabee, G. J., Rohling, M. L., Boone, K. B., Kirkwood, M. W., Schroeder, R. W., Suhr, J. A., & Conference Participants. (2021). American Academy of Clinical Neuropsychology (AACN) 2021 consensus statement on validity assessment: Update of the 2009 AACN consensus conference statement on neuropsychological assessment of effort, response bias, and malingering. *The Clinical Neuropsychologist*, *35*(6), 1053–1106. https://doi.org/10.1080/13854046.2021.1896036

Taylor, J. R. (2022). *An introduction to error analysis: The study of uncertainties in physical measurements* (3rd ed.). University Science Books.

Trahan, L. H., Stuebing, K. K., Fletcher, J. M., & Hiscock, M. (2014). The Flynn effect: A meta-analysis. *Psychological Bulletin*, *140*(5), 1332–1360. https://doi.org/10.1037/a0037173

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, *3*(2), 54–56. https://doi.org/10.1037/h0054962

VandenBos, G. R. (Ed.). (2007). *APA dictionary of psychology*. American Psychological Association.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213–217. https://doi.org/10.1037/0033-2909.83.2.213

Watkins, M. W., Canivez, G. L., Dombrowski, S. C., McGill, R. J., Pritchard, A. E., Holingue, C. B., & Jacobson, L. A. (2022). Long-term stability of Wechsler Intelligence Scale for Children–fifth edition scores in a clinical sample. *Applied Neuropsychology: Child*, *11*(3), 422–428. https://doi.org/10.1080/21622965.2021.1875827

Watkins, M. W., & Smith, L. G. (2013). Long-term stability of the Wechsler Intelligence Scale for Children–Fourth edition. *Psychological Assessment*, *25*(2), 477–483. https://doi.org/10.1037/a0031653

Watson, D. (2015). Intelligence testing. In E. A. Polloway (Ed.), *The death penalty and intellectual disability* (pp. 113–140). American Association on Intellectual and Developmental Disability.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale–Fourth Edition (WAIS-IV)*. Pearson.

Wechsler, D. (2024). *Wechsler Adult Intelligence Scale–Fifth Edition (WAIS-5)*. NCS Pearson.

Wechsler, D., Raiford, S. E., & Presnell, K. (2024). *Wechsler Adult Intelligence Scale: Technical and Interpretive Manual* (5th ed.). NCS Person.

Weiss, L. G., Gregoire, J., & Zhu, J. (2015). Flaws in Flynn effect research with the Wechsler Scales. *Journal of Psychoeducational Assessment*, *34*(5), 411–420. https://doi.org/10.1177/0734282915621222

Wendelborg, C., Garrels, V., Sigstad, H. M. H., & Dean, E. E. (2022). Recruitment and work arrangements for employees with intellectual disability in competitive employment. *Journal of Policy and Practice in Intellectual Disabilities*, *19*(4), 350–359. https://doi.org/10.1111/jppi.12418

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*(1), 23–40. https://doi.org/10.1007/BF02287917

Winter, E. L., Trudel, S. M., & Kaufman, A. S. (2024). Wait, Where's the Flynn effect on the WAIS-5? *Journal of Intelligence*, *12*(11), Article 118. https://doi.org/10.3390/jintelligence12110118

Wongupparaj, P., Wongupparaj, R., Morris, R. G., & Kumari, V. (2023). Seventy years, 1000 samples, and 300,000 SPM scores: A new meta-analysis of Flynn effect patterns. *Intelligence*, *98*, Article 101750. https://doi.org/10.1016/j.intell.2023.101750

Appendix A

Technical Terms Defined

Table A1
Explanations of Technical Terms

Term	Symbol	Meaning
Observed score	X	The score obtained by a person on a specific test.
True score	T	The average score a person would obtain on a specific test if it could be administered repeatedly without carryover effects under a variety of plausible testing circumstances and environments.
Reliability coefficient	ρ_{XX}	The ratio of true score variance to observed score variance (often estimated from the correlation of scores from a test given twice).
Population mean	μ	The mean score of a particular population on a specific test.
Population standard deviation	σ	The standard deviation of scores of a particular population on a specific test.
Standard error of measurement	σ_{T-X}	The standard deviation of the distance between true scores and observed scores.
Standard error of estimation	$\sigma_{T-\hat{T}}$	The standard deviation of the distance between true scores and estimated true scores.
Confidence interval	CI	A range of scores that has a specified probability of containing the true score.
Expected value	$\mathscr{E}()$	A function that returns the long-term average of a random variable.

Appendix B

Random Variables and Descriptive Statistics

A *random variable* has values determined by a random process (e.g., a coin flip). A person's score on an IQ test is the sum of many random variables, including an idiosyncratic shuffle and mixing of parental genes, the twists and turns of personal experience, and moment-to-moment fluctuations of measurement error.

The *expected value* operator $\epsilon()$ is a function that takes the long-term average of the value inside the parentheses. Thus, the expected value of the random variable X is the population mean (μ) (see Equation B1).

$$\epsilon(X) = \mu_X. \tag{B1}$$

The *deviation* of any particular value of X from its population mean is computed with Equation B2.

$$X - \mu_X. \tag{B2}$$

We would like to know the typical size of a variable's deviation from its mean. Taking the mean of all deviations is not helpful because the average deviation is always exactly 0 (see Equation B3).

$$\begin{aligned} \epsilon(X - \mu_X) &= \epsilon(X) - \epsilon(\mu_X) \\ &= \mu_X - \mu_X \\ &= 0. \end{aligned} \tag{B3}$$

To make all deviations positive, we square them first. The average squared deviation is a statistic called the *variance* (σ^2), as seen in Equation B4.

$$\sigma_X^2 = \epsilon((X - \mu_X)^2). \tag{B4}$$

The typical size of a deviation is a statistic called the *standard deviation* (σ), which is the square root of the variance (see Equation B5).

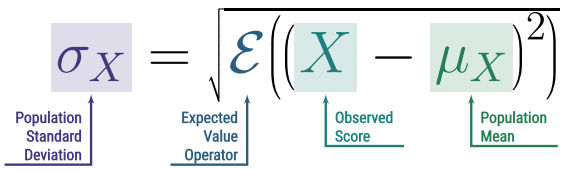
$$\sigma_X = \sqrt{\sigma_X^2}. \tag{B5}$$

Thus, the standard deviation is a special kind of average for deviations—the square root of the average squared deviation (see Figure B1).

(*Appendices continue*)

Figure B1

The Standard Deviation Is the Square Root of the Average Squared Deviation From the Population Mean



Note. The expected value operator ϵ is a function that returns the long-term average of a random variable.

It is useful to summarize how related two variables are. One such measure, the *covariance*, is the average product of two variables' deviations from their respective means (see Equation B6).

$$\sigma_{XY} = \epsilon((X - \mu_X)(Y - \mu_Y)). \tag{B6}$$

If X and Y tend to deviate from their respective means in the same direction, the product of their deviations will be large, on average. If their deviations are unrelated, the average product of their deviations will be zero. If two variables tend to deviate from their respective means in opposite directions, the average product of their deviations will be negative.

The covariance is called an *unstandardized* statistic because the values it produces do not have a fixed, standardized interpretation. A covariance of 10, for example, could represent a strong relationship between two variables or a very weak relationship, depending on the size of the variables' standard deviations.

To obtain a standardized measure of the relationship between variables, it can be useful to convert variables to have a common metric. To *standardize* a variable means to convert variables so that they have a mean of 0 and standard deviation of 1 using the z-score formula (see Equation B7).

$$z_X = \frac{X - \mu_X}{\sigma_X}. \tag{B7}$$

The *correlation* between two variables X and Y measures the strength of linear relationship between the two variables. The most commonly used correlation coefficient, the Pearson product-moment correlation (ρ), is the average product of their z-scores (see Equation B8).

$$\rho_{XY} = \epsilon(z_X z_Y). \tag{B8}$$

A correlation coefficient is a standardized covariance because it is the covariance of two standardized variables (z-scores). More importantly, the correlation coefficient is a standardized statistic because its interpretation is the same for any pair or variables. Correlations range from −1 to 1, with values near 0 indicating weak relationships and values near −1 or 1 indicating strong negative or positive relationships, respectively.

Appendix C

Reliability, Standard Error of Measurement, and Confidence Intervals

Standard Error of Measurement

Because true scores T and measurement errors e are uncorrelated, the variance of an observed score σ_X^2 is the sum of true score variance and the error variance, as shown in Figure C1.

From the equation in Figure C1, solving for σ_T^2, we see that $\sigma_T^2 = \sigma_X^2 - \sigma_e^2$. Therefore, substituting into the equation in Figure 5, and solving for σ_e, as seen in Equation C1.

$$\begin{aligned} \rho_{XX} &= \frac{\sigma_T^2}{\sigma_X^2} \\ \sigma_X^2 \rho_{XX} &= \sigma_T^2 \\ \sigma_X^2 \rho_{XX} &= \sigma_X^2 - \sigma_e^2 \\ \sigma_e^2 &= \sigma_X^2 - \sigma_X^2 \rho_{XX} \\ \sigma_e &= \sqrt{\sigma_X^2 - \sigma_X^2 \rho_{XX}} \\ \sigma_{X-T} = \sigma_e &= \sigma_X \sqrt{1 - \rho_{XX}} \end{aligned} \tag{C1}$$

Standard Error of Estimation

Both the standard error of measurement (SEM) and the standard error of estimation derive from the reliability coefficient. Whereas the SEM σ_{X-T} represents the typical distance from the observed score X to a true score T, the standard error of estimation is the typical distance from an estimated true score $\hat{T}$ to the actual true score T. Thus, in this context, the standard error of estimation can be symbolized as $\sigma_{\hat{T}-T}$.

The standard error of estimation is always smaller than the SEM because it is the SEM multiplied by the square root of the reliability coefficient, which is a value always less than 1 (see Equation C2).

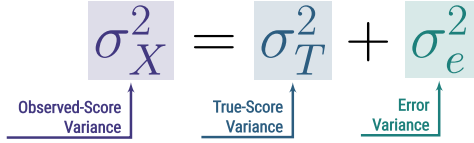
$$\begin{aligned} \sigma_{\hat{T}-T} &= \sigma_{X-T}\sqrt{\rho_{XX}} \\ &= \sigma_X \sqrt{\rho_{XX} - \rho_{XX}^2} \\ &= 15\sqrt{.97 - .97^2} \\ &\approx 2.56. \end{aligned} \tag{C2}$$

The 95% margin of error is obtained by multiplying the standard error of estimate by the z-score associated with the middle 95% of the normal distribution ($z_{95\%} \approx \pm 1.96$). Thus, the margin of error is computed with Equation C3.

$$95\% \text{ Margin of Error} = z_{95\%}\sigma_{\hat{T}-T}. \tag{C3}$$

Figure C1

Observed Score Variance Is the Sum of True Score Variance and Error Variance



(*Appendices continue*)

Estimated True Scores

In simple linear regression, the best linear equation for predicting one variable with another is seen in Figure C2.

Starting with the equation in Figure C2, we can substitute T for Y, remembering that $\mu_X = \mu_T$ and $\rho_{XX} = \frac{\sigma_T^2}{\sigma_X^2} = \rho_{XT}^2$ (see Equation C4):

$$\begin{aligned}\hat{T} &= \frac{\sigma_T}{\sigma_X}\rho_{XT}(X - \mu_X) + \mu_T \\ \hat{T} &= \sqrt{\frac{\sigma_T^2}{\sigma_X^2}}\sqrt{\rho_{XT}^2}(X - \mu_X) + \mu_X \\ \hat{T} &= \sqrt{\rho_{XX}}\sqrt{\rho_{XX}}(X - \mu_X) + \mu_X \\ \hat{T} &= \rho_{XX}(X - \mu_X) + \mu_X.\end{aligned} \tag{C4}$$

The estimated true score $\hat{T}$ of an IQ = 70 with a population mean of $\mu = 100$ and a reliability coefficient of $\rho_{XX} = .97$ is calculated via Equation C5:

$$\begin{aligned}\hat{T} &= \rho_{XX}(X - \mu_X) + \mu_X \\ 70.9 &= .97(70 - 100) + 100\end{aligned}. \tag{C5}$$

Confidence Intervals

The confidence interval is the combination of a point estimate ($\hat{T}$) and a two-tailed margin of error ($z\sigma_{\hat{T}-T}$). Combining Equations C2 and C5, the equation in Figure C3 gives the 95% confidence interval.

Inserting the values at hand, the 95% confidence interval for X is computed with Equation C6.

$$\begin{aligned}95\%\text{CI} &= \hat{T} \pm z_{95}\sigma_{\hat{T}-T} \\ &= \rho_{XX}(X - \mu_X) + \mu_X \pm z_{95}\sigma_X\sqrt{\rho_{XX} - \rho_{XX}^2} \\ &= .97(70 - 100) + 100 \pm 1.96 \cdot 15\sqrt{.97 - .97^2} \\ &\approx 70.9 \pm 1.96 \times 2.56 \\ &\approx 70.9 \pm 5.02 \\ &\approx [65.88, 75.92] \\ &\approx [66, 76].\end{aligned} \tag{C6}$$

Figure C2
Simple Linear Regression Equation in Which X Predicts Y

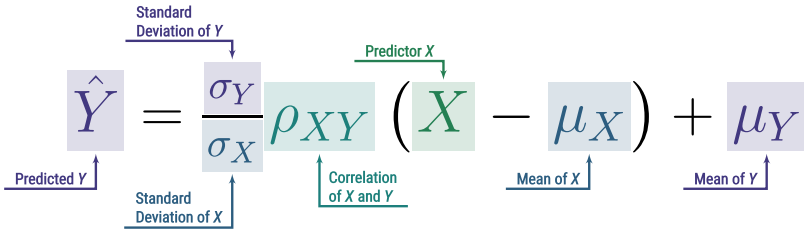
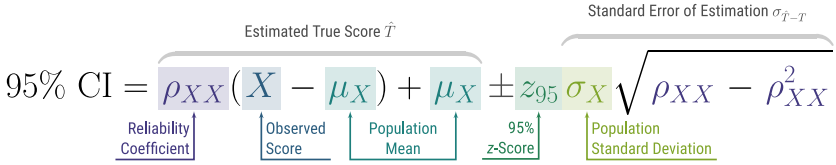


Figure C3
The 95% Confidence Interval of an Observed Score X



Received November 18, 2025
Revision received January 1, 2026
Accepted January 8, 2026 ■