APPLIED NEUROPSYCHOLOGY: CHILD, 5: 56–66, 2016

ISSN: 2162-2965 print/2162-2973 online
DOI: 10.1080/21622965.2014.976863

Routledge
Taylor & Francis Group

Comparing Test Scores Using Information From Criterion-Related Validity Studies

A. Alexander Beaujean
Department of Educational Psychology, Baylor University, Waco, Texas

Sean M. McGlaughlin
Deer Valley Unified School District, Phoenix, Arizona

There is frequently a need to compare a client's test scores from different instruments. If the scores come from instruments that use the same scale, it is tempting to compare the scores directly. Unfortunately, this method can lead clinicians to believe that there is a large difference between scores when the difference is minimal. As an alternative, we outline a method for score comparison that uses information from criterion-related validity studies. Using three examples, we show why this method is more psychometrically sound, produces more accurate comparison scores, and requires little extra work for clinicians than the direct comparison approach. To make the score comparison process easy for clinicians to use, we include an appendix that demonstrates how to implement this method in Microsoft Excel and the free **R** program.

Key words: *psychological assessment, score comparison, score exchangeability, test scores*

Psychological assessment is an important part of clinical practice (Castillo, Curtis, & Gelley, 2012; Evers et al., 2012; Norcross & Karpiak, 2012). Although its use for diagnosis is still important (Swets, Dawes, & Monahan, 2000), assessment sequelae extend beyond this. Psychological assessment is also used for treatment (Lambert & Vermeersch, 2013), determining service eligibility (Swenson, 2013), and even death penalty decisions (Gresham, 2009).

Yet, despite its importance, typical training in assessment is rather minimal (Aiken, West, & Millsap, 2008; Haverkamp, 2013; Stedman, Hatch, & Schoenfeld, 2001). Even worse, the training that many receive in core psychometrics—the foundation for properly interpreting assessment results—is cursory (Childs & Eyde, 2002; Handler & Smith, 2012). Consequently, it is not surprising that internship sites find their trainees underprepared to conduct psychological assessments (Clemence & Handler, 2001), psychologists often select training in assessment-related skills for their continuing education (Neimeyer, Taylor, & Philip, 2010), manuscripts are frequently rejected because of errors stemming from a lack of formal training in measurement (Reynolds, 2008), and experts suggest probing psychologists' psychometric knowledge in challenging courtroom testimony (Reynolds & Milam, 2011).

One of the basic assessment competencies for professional psychological practice is the ability to interpret and integrate scores from multiple psychological tests (Fouad et al., 2009). A key skill in this competency is the ability to compare test scores. This skill is important as it is used to compare results from different assessments (i.e., compare current results to previous results), as well as to compare scores from different tests within the same assessment (e.g., cognitive ability and academic achievement). The methods that psychologists use to compare scores, however, are not always aligned with best practice.

Address correspondence to A. Alexander Beaujean, Department of Educational Psychology, Baylor University, One Bear Place #97301, Waco, TX 76798-7301. E-mail: Alex_Beaujean@baylor.edu

TYPICAL METHODS FOR SCORE COMPARISONS

Direct Comparison

Modern psychological tests frequently place scores on well-known scales, such as the IQ scale or the *T* scale. The use of common scales makes it easy to believe that scores from two tests (i.e., $Test_1$ and $Test_2$) can be directly compared. As some authors advocate (e.g., Flanagan, Ortiz, & Alfonso, 2007), if $Test_1$ and $Test_2$ ostensibly have the same mean and standard deviation (SD), then it is easy to believe that an individual's score on $Test_1$ should be the same as the individual's score on $Test_2$.

The direct comparison method makes two major assumptions: The score variability for $Test_1$ is the same as that for $Test_2$, and the score mean of $Test_1$ is the same as that for $Test_2$. Because test publishers purposefully give their tests' scores a common mean and SD, it is easy to think these assumptions are appropriate. We argue that this assumption is false: Test scores from different instruments are typically not measured on equivalent (i.e., directly comparable) units. We present two lines of support for this argument. The first revolves around how test scores are usually created and the second involves examining empirical evidence.

The original scores obtained from a test are *raw scores* and are calculated by quantifying a respondent's performance, such as summing the items answered correctly. Raw scores are problematic because their units (e.g., number of items answered correctly) are not interpretable by those not very familiar with the test (Angoff, 1971). Consequently, most test developers transform the raw scores to values that are easier to interpret. A common score transformation is the *Z* score transformation.[1] The formula to transform raw scores to *Z* scores is

$$Z\,\text{Score} = \frac{\text{Raw Score} - \text{Raw Score Mean}}{\text{Raw Score Standard Deviation}} \qquad (1)$$

where the raw score is the raw test value from an examinee selected from the normative sample, and the raw score mean and raw score standard deviation are the mean and SD of all the raw scores, respectively, produced by the normative sample.

Equation 1 shows that the *Z*-score transformation requires two steps. First, it requires subtracting the normative sample's average raw score from each respondent's raw score, the results of which are sometimes called a *mean–deviation score*. Doing this makes the average *Z* score equal to 0 and changes the score's interpretation. While raw scores measure some aspect of the respondents' actual performance on the particular test, *Z* scores measure the difference between the respondents' raw scores and the normative sample's average score. Second, the *Z*-score transformation requires dividing the mean–deviation score by the normative sample's SD, which has two major effects on the resulting scores. First, it converts the unit from that used by the raw scores (e.g., number of items answered correctly) to an SD. Second, it makes the SD of the *Z* scores equal 1.

Because *Z* scores require the use of decimals and half of a sample's *Z* scores will always be negative, test publishers often alter the mean and SD to make the values non-negative integers. To convert the *Z* scores' scale, multiply them by the new SD of interest. Likewise, to convert the *Z* scores' mean, add the mean of interest to each score. For example, placing *Z* scores onto the IQ scale requires multiplying all the *Z* scores by 15, adding 100, and rounding the values to the nearest integer (Seashore, 1955).

An important part of our description for creating *Z* scores was their absolute dependence on the normative sample used to create them. This translates into practice as meaning that the same raw score will be converted to different *Z* scores depending on the sample used for their creation. If the test developers use representative samples, the difference for a given raw score's *Z* score between one sample and another should be small, but it will likely not be 0. Thus, because psychological tests use different normative samples to develop their scores, the resulting *Z* scores (and any subsequent scale transformations) are typically not directly comparable. Moreover, if the norming process was done multiple years apart, the Flynn effect (i.e., secular gains in average test scores of measures of cognitive ability; Flynn, 2012) could also create score differences.

To bolster the argument against using direct comparisons further, one need only examine the criterion-related validity (CRV) studies of most psychological tests. Such studies are usually conducted by having the *same* individuals take two tests (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999). When reports of these studies contain the means and SDs of the test scores, they often show that these values are not equivalent for the two tests. We demonstrate this in Table 1, which contains the means and SDs from CRV studies of some popular tests of cognitive ability. Thus, tests' CRV studies provide further evidence that the equal variability and equal means assumptions on which the direct comparison method relies are usually not met.

[1]This is an oversimplification of the transformation process, as techniques such as smoothing, continuous norming, and score normalization are also frequently employed during the raw-score conversion process.

TABLE 1
Sample of Criterion-Related Validity Studies

Test 1			*Test 2*			
Test	*Mean*	*SD*	*Test*	*Mean*	*SD*	*Reference*
RIAS	100.32	16.18	WISC-III	107.89	17.97	Reynolds & Kamphaus (2003)
WAIS-III	102.90	14.90	WAIS-IV	100.00	15.20	Wechsler (2008)
WPPSI-IV	100.20	12.70	WISC-IV	105.30	12.50	Wechsler, Coalson, & Raiford (2012)
WAIS-III	107.00	18.70	SB-V	101.50	14.40	Roid (2003)

Note. Reported scores are the Full-Scale IQ or its equivalent. RIAS = Reynolds Intellectual Assessment Scales; WAIS-III = Wechsler Adult Intelligence Scale-Third Edition; WAIS-IV = Wechsler Adult Intelligence Scale-Fourth Edition; WISC-III = Wechsler Intelligence Scale for Children-Third Edition; WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition; WPPSI-IV = Wechsler Preschool and Primary Scale of Intelligence-Fourth Edition; SB-V = Stanford-Binet Intelligence Scales-Fifth Edition.

Correlation Method

The correlation method uses the correlation between the scores on $Test_1$ and $Test_2$ to predict the score on $Test_2$ from the score on $Test_1$ (Schneider, 2013). One benefit of this method is that it accounts for regression to the mean for extreme scores (for an accessible description of the phenomenon of the regression to the mean, see Healy & Goldstein, 1978). For extremely low scores on $Test_1$, the correlation method will predict scores on $Test_2$ that are not as low—that is, $Test_2$'s predicted score will be closer to its mean than $Test_1$'s score is to its mean. The same applies to extremely high scores as well.

Although the correlation method is somewhat better than direct comparisons, it also has some problems. First, it makes the same assumptions of the direct comparison method. Namely, because $Test_1$ and $Test_2$ were transformed to have the same mean and SD within their own standardization sample, these means and SDs are equivalent across instruments. Thus, the same arguments against using direct comparisons can also be used against using the correlation method. Second, this method assumes that if $Test_1$ and $Test_2$ use the same scale, they are interchangeable. Consequently, $Test_1$ will predict the same score for $Test_2$ as $Test_2$ predicts for $Test_1$. This second assumption will become more evident in the second example that we present later in this manuscript. Third, this method does not account for the Flynn effect, which could change the average score values if the tests were normed at different times.

As an alternative to the direct comparison and correlation methods, we propose a third method for score comparison. Our method is very similar to the correlation method, but it uses the means, SDs, and correlations from CRV studies to predict test scores.

USING CRITERION-RELATED VALIDITY STUDIES FOR SCORE COMPARISON

Using information from CRV studies to compare scores requires the use of simple regression. One of the major purposes of simple regression is to predict the value on one variable from a value on another variable (Cohen, Cohen, West, & Aiken, 2003). Regression may be unfamiliar to, or not well understood by, some clinicians. Consequently, we review the basic ideas.

Simple Regression

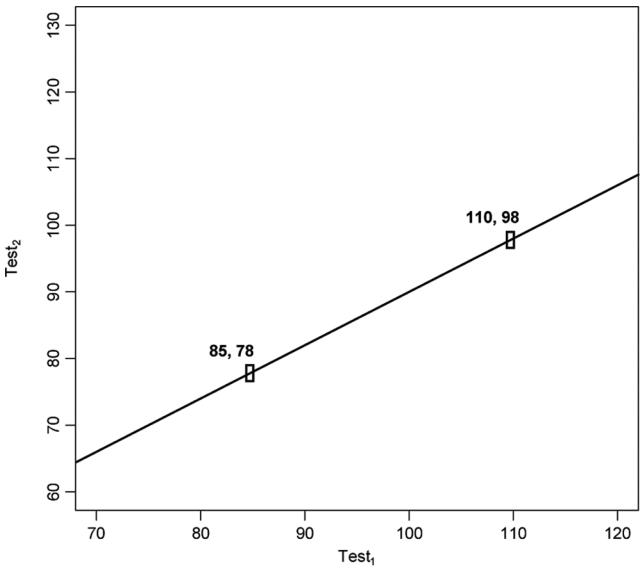
Understanding how to use simple regression to predict test scores requires little more than knowledge about linear functions. An example of a linear function is

$$\underset{\text{outcome variable}}{Test_2} = \underset{\text{intercept}}{a} + \underset{\text{slope}}{b} \times \underset{\text{predictor variable}}{Test_1} \quad (2)$$

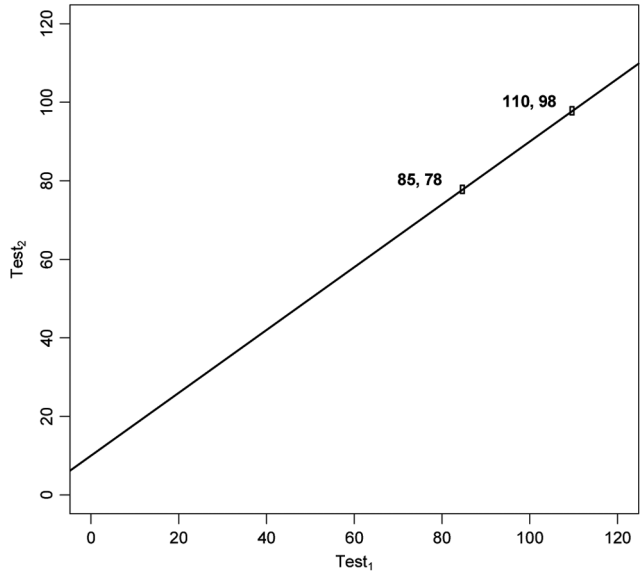
The two major parts of Equation 2 are the b and a terms. The b term is the *slope* and represents the "rise" of a line over its "run." For our purposes, b indicates how much $Test_2$'s scores change when $Test_1$'s scores change one unit. The a term in Equation 2 is the *intercept*. It represents the value of $Test_2$ when $Test_1$ has a value of 0. As 0 is not a possible value for many psychological tests, another way of thinking about a is that it is a constant that accounts for the difference between the observed mean of $Test_2$ and the mean of $Test_2$ predicted by $Test_1$ when it is multiplied by b. If there are no mean differences between $Test_1$ and $Test_2$, then the intercept will be close to 0.

An example of a linear function is shown in Figure 1a. In the figure, the slope is .80, meaning that when $Test_1$ increases by 1 unit, then $Test_2$ is predicted to increase 0.80 units. The intercept is 10, meaning that individuals who obtained a score of 0 on $Test_1$ are predicted to earn a score of 10 on $Test_2$. Figure 1b shows a picture of Figure 1a zoomed out to show the intercept.

After finding the values for a and b, Equation 2 can be used to predict scores on $Test_2$ from $Test_1$ scores. The predicted $Test_2$ values are symbolized as $\widehat{Test_2}$ and read as $Test_2$-hat. The predicted values for $Test_2$ are



(a) Plot of obtained test scores.



(b) Plot zoomed out to show intercept.

FIGURE 1 Plot of a simple linear function between two test scores on the IQ scale.

calculated as

$$\widehat{Test_2} = a + b \times Test_1 \tag{3}$$

Accounting for Imperfect Relationships Between Tests

Very seldom are there perfect relationships in psychological assessment. Instead, most work deals with imperfect probabilistic relationships that involve *error*. Accounting for this error requires a slight alteration of Equation 2 to include an error term:

$$\underset{\text{variable}}{Test_2\ \text{outcome}} = \underset{\text{intercept}}{a} + \underset{\text{slope}}{b} \times \underset{\text{variable}}{Test_1\ \text{predictor}} + \underset{\text{error}}{e} \tag{4}$$

Combining Equations 3 and 4 shows the error term is just the difference between the actual value of $Test_2$ and the predicted value of $Test_2$ (i.e., $Test_2 - \widehat{Test_2}$). Consequently, the best values to use for a and b are those that make the error term as small as possible, which then make $Test_2$ and $\widehat{Test_2}$ as close as possible.

The interpretation of the values in Equation 4 is very similar to the interpretation of values in Equation 2, except that now probabilistic language is used. The slope, b, is how much $Test_2$ is expected (i.e., on average) to change when $Test_1$ increases by 1 point. The intercept, a, is the expected value of $Test_2$ when $Test_1$ is 0. The predicted $Test_2$ score, $\widehat{Test_2}$, is still calculated using Equation 3, but its interpretation is now the expected value of $Test_2$ for all the respondents who earned a specific score on $Test_1$.

Relationship Between Regression Slope and Correlation

Regression and correlation are intimately related to each other. To understand this, imagine transforming the scores from $Test_1$ and $Test_2$ to Z scores (see Equation 1). Then, use these Z scores to estimate the values for the slope and intercept (i.e., a and b) in Equation 4. The use of Z scores in this situation causes two things occur: (a) The intercept becomes 0; and (b) the slope becomes the Person correlation coefficient.[2]

As we stated previously, using correlations to predict test scores is problematic. CRV studies, however, usually opt to only report test scores' correlations instead of slopes. Fortunately, converting a correlation to a slope for a simple regression is a straightforward calculation:

$$b = r_{12} \times \frac{SD_2}{SD_1} \tag{5}$$

where r_{12} is the correlation between $Test_1$ and $Test_2$, and SD_1 and SD_2 are the SDs for $Test_1$ and $Test_2$, respectively (Cohen et al., 2003). To use Equation 5, the data for all three statistics (i.e., r_{12}, SD_1, and SD_2) should all come from the same sample.

[2]In some texts, the slope estimated using Z scores is called a *standardized regression coefficient*.

PREDICTING $TEST_2$ SCORES FROM $TEST_1$ SCORES

Clinicians are seldom given the values of a and b that would allow them to predict test scores. Nonetheless, we can still predict the score on $Test_2$ from a score on $Test_1$. To do so requires five values: (a) mean of $Test_2$ (M_2); (b) SD of $Test_2$ (SD_2); (c) mean of $Test_1$ (M_1); (d) SD of $Test_1$ (SD_1); and (e) the correlation between $Test_2$ and $Test_1$ (r_{12}). Such values are frequently provided in CRV studies. To use these values, modify Equation 3 to take into account the correlation–slope relationship given in Equation 5:

$$\underset{\text{predicted test score}}{\widehat{Test_2}} = \underset{\text{intercept}}{M_2} + \underbrace{r_{12} \times \frac{SD_2}{SD_1}}_{\text{slope}} \times \underbrace{(Test_1 - M_1)}_{\text{collected test score}} \quad (6)$$

Confidence Intervals for Predicted Scores

When using test scores, it is typically better to interpret the confidence interval (CI) around the score than the score itself, as including a CI gives an indication of the precision of the test score (AERA, APA, & NCME, 1999; Hall v. Florida, 2014). The same logic applies when predicting test scores; we want both the predicted value and the CI for the predicted value.

There are three components involved in creating a CI for a predicted test score: (a) the predicted test score ($\widehat{Test_2}$); (b) the desired confidence level, which will be 1 minus the type 1 error rate, *alpha* (i.e., $1 - alpha$); and (c) the amount of error in the regression equation used to predict the test score ($\sigma_{\widehat{Test_2}}$). Combing those three parts produces a $(1 - alpha)$% CI:

$$\widehat{Test_2} \pm \sigma_{\widehat{Test_2}} \times N_{alpha/2} \quad (7)$$

where $Nalpha/2$ is the value from a standard normal distribution (i.e., a normal distribution with a mean of 0 and SD of 1) that has $alpha/2$% of the area to the left and $alpha/2$% of the area to the right.

For relatively large sample sizes, the value for $\sigma_{\widehat{Test_2}}$ can be approximated by

$$\sigma_{\widehat{Test_2}} \approx SD_2 \times \sqrt{1 - r_{12}^2} \quad (8)$$

where SD_2 is the SD of $Test_2$, and r_{12}^2 is the squared correlation between $Test_1$ and $Test_2$ (Crocker & Algina, 1986). Notice in Equation 8 that the size of $\sigma_{\widehat{Test_2}}$, and consequently the size of the CI, is directly related to the correlation between the two tests. Larger correlations between the two tests produce narrower CI widths.

EFFECTS OF USING THE INCORRECT MEANS AND STANDARD DEVIATIONS

What happens when incorrect mean or SD values are used to predict test scores? If only the means are wrong, then Equation 6 tells shows that the effect will be to overpredict or underpredict the value for $Test_2$ in a systematic fashion. For example, if there was a 10-point difference between the mean used in Equation 6 and the value it should be, then the predicted values for $Test_2$ will be overpredicted or underpredicted by 10 points.

The effect of using the wrong SDs is a little more complicated. According to Equation 5, the effect depends on the way the SDs are wrong. If the SD of *Test*2 is less than the SD of $Test_1$, then the predicted values move away from their mean value. Thus, $Test_2$ values are underpredicted when $Test_1$ scores are below the mean and overpredicted when $Test_1$ scores are above the mean. If the SD of $Test_2$ is greater than the SD of $Test_1$, then the predicted values for $Test_2$ move close to the mean. Thus, $Test_2$ scores will be overpredicted when $Test_1$ scores are below the mean and underpredicted when $Test_1$ scores are above the mean. If both the means and the variances are wrong, then the discrepancy between the correctly and incorrectly predicted values can be difficult to understand. To make it easier, we present an example that shows the effects of using incorrect means and SDs.

Example 1

In Table 2, we show some predicted scores for $Test_2$ based on $Test_1$ scores. The mean and SD for $Test_1$ are 100 and 15, respectively, and the obtained scores on $Test_1$ are either 100, 70, or 130 (representing the mean and values 2 SDs above and below the mean). The SDs for $Test_2$ are either the same as $Test_1$ (i.e., $SD = 15$), lower than $Test_1$ (i.e., $SD = 10$), or higher than $Test_1$ (i.e., $SD = 15$). The correlations between $Test_1$ and $Test_2$ are set at .70 and .90.

The first three comparisons in Table 2 contain the situation where the means and SDs are the same for both tests. We refer to this situation as the *baseline*. With a correlation of .70 or .90, the predicted values for $Test_2$ are relatively close to the $Test_1$ scores, although the predicted values using the .70 correlation are closer toward $Test_2$'s mean than the predicted values using .90. Another thing to notice about these comparisons is that the 95% CIs are much wider for the .70 correlation than the .90 correlation. This situation repeats itself for all comparisons in Table 2.

For the remaining comparisons (4–27), the assumptions of the means or SDs being the same for $Test_1$ and $Test_2$ are not met. For Comparisons 4 through 9, the SDs are the same for both tests, but the means differ. In Comparisons 4 through 6, the mean for $Test_2$ is 10

TABLE 2
Sample Predicted Values for $Test_2$

		$Test_2$									
				Correlation = .90				*Correlation = .70*			
Comparison	$Test_1$	M_2	SD_2	$Pred_2$	LB_{95}	UB_{95}	*Off*	$Pred_2$	LB_{95}	UB_{95}	*Off*
1	70	100	15	73	60	86	0	79	58	100	0
2	100	100	15	100	87	113	0	100	79	121	0
3	130	100	15	127	114	140	0	121	100	142	0
4	70	90	15	63	50	76	−10	69	48	90	−10
5	100	90	15	90	77	103	−10	90	69	111	−10
6	130	90	15	117	104	130	−10	111	90	132	−10
7	70	110	15	83	70	96	10	89	68	110	10
8	100	110	15	110	97	123	10	110	89	131	10
9	130	110	15	137	124	150	10	131	110	152	10
10	70	100	10	82	73	91	9	86	72	100	7
11	100	100	10	100	91	109	0	100	86	114	0
12	130	100	10	118	109	127	−9	114	100	128	−7
13	70	90	10	72	63	81	−1	76	62	90	−3
14	100	90	10	90	81	99	−10	90	76	104	−10
15	130	90	10	108	99	117	−19	104	90	118	−17
16	70	110	10	92	83	101	19	96	82	110	17
17	100	110	10	110	101	119	10	110	96	124	10
18	130	110	10	128	119	137	1	124	110	138	3
19	70	100	20	64	47	81	−9	72	44	100	−7
20	100	100	20	100	83	117	0	100	72	128	0
21	130	100	20	136	119	153	9	128	100	156	7
22	70	90	20	54	37	71	−19	62	34	90	−17
23	100	90	20	90	73	107	−10	90	62	118	−10
24	130	90	20	126	109	143	−1	118	90	146	−3
25	70	110	20	74	57	91	1	82	54	110	3
26	100	110	20	110	93	127	10	110	82	138	10
27	130	110	20	146	129	163	19	138	110	166	17

Note. The mean and standard deviation (SD) of $Test_1$ are 100 and 15, respectively. M_2 = mean of $Test_2$; SD_2 = SD of $Test_2$; $Pred_2$ = predicted score on $Test_2$ given $Test_1$ score; LB_{95} = lower bound of 95% confidence interval (CI); UB_{95} = upper bound of 95% CI; Off = how much the baseline predicted values (Comparisons 1 through 3) differ from the actual predicted value.

points below that of $Test_1$. Consequently, the predicted $Test_2$ scores are 10 points lower than those from their baseline scores. In Comparisons 7–9, the mean for $Test_2$ is 10 points greater than that of $Test_1$, making the predicted scores on $Test_2$ 10 points higher than their baseline scores. These underpredictions and overpredictions apply to both the correlations of .70 and .90.

Comparisons 10 through 18 repeat Comparisons 1 through 9, only add that the SD for $Test_2$ is one third smaller than that of $Test_1$ (i.e., the SD is 10 instead of 15). For Comparisons 10 through 12, there are no mean differences. For an average score on $Test_1$ (Comparison 11), the predicted value for $Test_2$ is the same as when the SD for both tests is the same (i.e., Comparison 2). This applies to both correlations. When the score on $Test_1$ deviates from its mean, however, the predicted values for $Test_2$ begin to move toward their mean. In Comparison 10, the score for $Test_1$ is 70. With a correlation of .90, the predicted value for $Test_2$ is 82 (18 points away from 100), while in Comparison 1, the predicted value is 73 (27 points away from 100). The difference between the predicted values is 9 points. With a correlation of .70, the predicted value is 86 (14 points away from 100), while in Comparison 1, the predicted value is 79 (21 points away from 100). The difference between the predicted values is 7 points. Similar phenomena occur for Comparison 12.

Comparisons 13 through 15 not only have an SD for $Test_2$ that is one third smaller than that for $Test_1$, but also add that the mean for $Test_2$ is 10 points lower than that of $Test_1$. For an average score on $Test_1$ (i.e., Comparison 14), the predicted value of $Test_2$ is 10 points lower than the score predicted from Comparison 2—the same phenomenon that occurred with Comparisons 4 through 6. This effect occurred with both correlations. When the score on $Test_1$ deviates from its mean, however, the predicted values for $Test_2$ move toward their means. For example, Comparison 13 has a score of 70 on $Test_1$. With a correlation of .90, the predicted value is 72 (18 points away from 90), while in Comparison 1, the predicted value is 73 (27 points away from 100). The difference between the predicted values is 1

point. With a correlation of .70, the predicted value is 76 (14 points away from 90), while in Comparison 1, the predicted value is 79 (21 points away from 100). The difference between the predicted values is 3 points.

In Comparison 15, the score on $Test_1$ is 130. With a correlation of .90, the predicted value for $Test_2$ is 108 (18 points away from 90), while in Comparison 3, the predicted value is 127 (27 points away from 100). The difference between the predicted values is 19 points. With a correlation of .70, the predicted value is 104 (14 points away from 90), while in Comparison 3, the predicted value is 121 (21 points away from 100). The difference between the predicted values is 17 points. Similar phenomena occur for Comparisons 16 through 18 as it did for Comparisons 13 through 15, only in the opposite direction: Low scores on $Test_1$ produce larger differences from the baseline than do high scores.

Integrating the results from Comparisons 13 through 18 shows that when the values of $Test_1$ are close to the actual mean of $Test_2$ and the actual SD of $Test_2$ is smaller than that of $Test_1$, the amount the predicted scores will be off by not accounting for the correct mean and the SD will be small. As the values for $Test_1$ depart from the actual mean of $Test_2$, then the amount of the predicted values for $Test_2$ will be off by not accounting for the correct mean and the SD begins to increase, reaching 19 points in some instances (e.g., Comparisons 15–16).

Comparisons 22 through 27 produce results that are the direct opposite of Comparisons 13 through 18. Here the SD for $Test_2$ is one third larger than that for $Test_1$. Consequently, the results in Comparisons 22 through 27 show the opposite effect of Comparisons 13 through 18. When the value of $Test_1$ is far from the actual mean of $Test_2$, then the amount of the predicted values will be off by not accounting for the correct mean and the SD will be small. As the values for $Test_1$ get closer to the actual mean of $Test_2$, then the amount of the predicted values for $Test_2$ will be off by not accounting for the correct mean and the SD begins to increase.

Example 2

As a different example, we use some of the CRV data from Table 1. In this scenario, a psychologist working in a hospital has a client who was administered the Third Edition of the Wechsler Intelligence Scale for Children (WISC-III; Wechsler, 1991) a few years before and earned a Full-Scale IQ (FSIQ) score of 65. The client's physician requested a new psychological evaluation, and as part of the test battery, the psychologist administered the Reynolds Intellectual Assessment Scales (RIAS; Reynolds & Kamphaus, 2003). What is the client's predicted composite index (CIX) score on the RIAS, given the FSIQ score on the WISC-III?

All the information needed to calculate the score is provided in Table 1 except the correlation, which the RIAS technical manual reports to be .76 (Reynolds & Kamphaus, 2003, p. 105). Plugging the known values into Equation 6 gives:

$$70.97 = 100.32 + 0.76 \times \left(\frac{16.18}{17.97}\right) \times (65 - 107.89)$$

Thus, the client's predicted CIX score on the RIAS is 71. To make a 95% CI around the predicted value, plug the known values into Equations 8 and 7:

$$\sigma_{\widehat{RIAS}} \approx 16.18 \times \sqrt{1 - .76^2} = 10.52$$

and

$$70.97 \pm 10.52 \times 1.96$$

Thus, the 95% CI is 50 and 92. Under the assumption that the mean and SD for both instruments were 100 and 15, respectively, the psychologist would have calculated a predicted RIAS CIX score of 73 (95% CI [54, 93]). The Appendix shows how to calculate predicted scores and their CIs use CRV data in **R** and Microsoft Excel.

Now, change the scenario slightly so that the client had previously earned a RIAS CIX score of 65 and the psychologist wanted to predict the WISC-III FSIQ score. Plugging the known values into Equation 6 returns:

$$78.08 = 107.89 + 0.76 \times \left(\frac{17.97}{16.18}\right) \times (65 - 100.32)$$

Thus, the predicted WISC-III FSIQ score is 78. Using Equations 8 and 7 to make the 95% CI around the predicted value produces:

$$\sigma_{\widehat{WISC-III}} \approx 17.97 \times \sqrt{1 - .76^2} = 11.68$$

and

$$78.08 \pm 11.68 \times 1.96$$

making the 95% CI 55 and 101.

Under the assumption that the mean and SD for both instruments were 100 and 15, respectively, the psychologist would have a predicted WISC-III FSIQ of 73 (95% CI [54, 93]). These are the exact same values predicted for the RIAS CIX from the WISC-III when not accounting for the mean and SD differences. Thus, as we previously stated, a product of the correlation method is that the predicted values for $Test_1$ based on $Test_2$ will be exactly the same as the predicted values for $Test_2$ based on $Test_1$.

TABLE 3
Random Sample of 15 Students' Scores on the Reynolds Intellectual Assessment Scales and Wechsler Intelligence Scale for Children-Third Edition

	Predicted Score			Actual–Predicted Score Difference		
RIAS CIX	*WISC-III FSIQ*[a]	*Correlation*	*CRV*	*WISC-III FSIQ*[a]	*Correlation*	*CRV*
57	70	77	74	13	20	17
79	77	83	79	−2	4	0
80	60	70	68	−20	−10	−12
86	81	86	82	−5	−0	−4
87	86	89	85	−1	2	−2
89	73	79	76	−16	−10	−13
89	97	98	93	8	9	4
90	69	76	74	−21	−14	−16
95	95	96	91	0	1	−4
97	106	105	99	9	8	2
98	112	109	103	14	11	5
100	102	102	96	2	2	−4
101	118	114	107	17	13	6
114	121	116	109	7	2	−5
119	116	112	106	−3	−7	−13

RIAS = Reynolds Intellectual Assessment Scales; CIX = composite index score; WISC-III = Wechsler Intelligence Scale for Children-Third Edition; FSIQ = Full-Scale IQ; CRV = criterion-related validity; difference = difference between predicted RIAS value and actual RIAS value.

[a]This is the *direct comparison* method as it compares the WISC- III and the RIAS scores directly.

This situation highlights the fact that no matter what test is used to predict the other, there is no reason to believe that clients who earn a given score on $Test_1$ will produce the exact the same score on $Test_2$. Even if they gave the exact same performance on both tests, clients will usually not produce the same scores because the tests' metrics are not equivalent. Moreover, this example echoes the results shown in Example 1: Even though a correlation of .76 seems strong, the CIs around the predicted test scores are large and indicate the lack of precision involved in predicting the RIAS from the WISC-III (and vice versa).

Example 3

For the third example, we compared the CRV, correlation, and direct comparison methods using a data set containing scores from 206 students referred for special education services in a large, public Midwestern school district. All students had WISC-III FSIQ and the RIAS CIX scores. Values for a random sample of 15 students are shown in Table 3.

To contrast the different score comparison methods, we used the methods given by Bland and Altman (1999), which are shown in Table 4. First, we estimated the methods' *bias*, which is the average amount of difference between the predicted RIAS score and the actual RIAS score. The correlation method has almost no bias for this sample, while the CRV and direct comparison methods both underpredict RIAS CIX scores by approximately 3.8 points. Bias is not a major problem in predicting test scores, however, as the effect is systematic and can be corrected. For the CRV and direct comparison methods, this would involve adding approximately 3.8 points to each predicted RIAS CIX score.

TABLE 4
Results From Contrasting the Different Score Comparison Methods

	Score Comparison Method		
Measure	*Direct*	*Correlation*	*Criterion-Related Validity*
Bias	−3.84	−0.01	−3.88
95% Limits of Agreement (range)	−23.61 - 15.92	−16.4 - 16.38	−19.74 - 11.98
95% Limits of Agreement Width	39.54	32.78	31.71
Bias Standard Error	1.20	1.00	0.96

Note. Values were estimated using 206 students with composite index scores on the Reynolds Intellectual Assessment Scales and Full-Scale IQ scores on the Wechsler Intelligence Scale for Children-Third Edition.

A more useful measure to compare the methods is the range of differences between the predicted and actual scores. This is captured by the *limits of agreement* (LoA), which define the range within which most differences between the predicted and actual RIAS CIX scores will lie. In other words, LoA provide a range of how discrepant the predicted RIAS CIX values are from the actual values. We expect that most of the differences would be contained in the 95% LoA, so that is what we show in Table 4. Following Bland and Altman (1986), interpretation of the 95% LoA for the CRV method is: The predicted RIAS CIX scores produced by using the WISC-III FSIQ and Reynolds and Kamphaus's (2003) WISC-III CRV study may be 20 points below or 12 points above the actual RIAS CIX scores. Similar interpretations follow for the direct comparison and correlation methods' 95% LoA. Whether this amount of lack of agreement is acceptable or unacceptable is a decision for those wishing to compare the scores.

As a point of comparison, the width of the 95% LoA for the three different score comparison methods shows that the CRV method produced the smallest LoA, with the correlation method having a slightly larger width of approximately 1 point. The direct comparison method, however, has a much wider LoA. This indicates that on average, in this sample, the direct comparison method is much less accurate than the CRV or correlation method.

The last value in Table 4 is the bias value's standard error. It is a measure of the precision of the bias and 95% LoA values. The CRV study method produced the most precise estimates, with the correlation method being slightly less precise and the direct comparison method producing the least precise estimates.

RELIABILITY EFFECTS

Throughout this manuscript, we have assumed that the test scores are perfectly reliable. This is seldom the case with psychological measures (Shrout, 1998). Unreliability in either $Test_1$ or $Test_2$ makes their calculated correlation artificially small. This affects the predicted scores in two ways. First, it forces the predicted values of $Test_2$ to be closer to their mean. To see this, compare the results in Table 2 using the correlation of .70 versus using the correlation of .90. In the extreme case of having no score reliability, the regression slope would be 0 and every predicted value for $Test_2$ would be the mean of $Test_2$. The second way unreliability, and its subsequent smaller correlations, affects score prediction is through Equation 8. Artificially small correlations produce regression error estimates that are too large. As this error value is used in Equation 7, it has the additional effect of widening the CIs.

There are ways to correct Equation 6 and Equation 7 to account for unreliability in test scores, but their complexity places them beyond the scope of the current article. Interested readers should consult Charter (1996) and Ree and Carretta (2006) for more information.

IMPLICATIONS FOR PRACTICE

In this manuscript, we discussed three methods of comparing scores and showed why the method using CRV studies is usually the best one to use. The implications for professional practice are fourfold.

First, unless authors of a test have specifically designed the scores to be directly comparable to those from another test, scores from different tests should not be directly compared to each other quantitatively. To make a test's scores be directly comparable with scores from another test requires either the same norming sample to be used for both instruments or some advanced psychometric techniques such as score equating (Dorans, 2004). Currently, this typically only happens with large-scale standardized tests that produce multiple equivalent forms of the test.

This prohibition on quantitative comparisons does not preclude *qualitative* comparisons. For example, if a client earned a score of 120 on $Test_1$ and 125 on $Test_2$, both of which were on the IQ scale, then a statement such as the following would be appropriate:

> On $Test_1$ and $Test_2$ the client's performance was above the average performance of the norming samples on both tests.

Problems only start to arise when making statements such as, "The client scored 5 points better on $Test_2$ than $Test_1$" (a direct score comparison) without first placing the test scores on comparable metrics.

Second, to be able to compare scores correctly requires CRV studies. These are often found in the test's technical manual or peer-reviewed articles. In the best-case scenario, these studies would be done with a strong research design, such as using random sampling and counterbalancing the administered tests. While such sampling may be used in CRV articles, this seldom occurs in the CRV studies reported in technical manuals. Typically, those studies have samples that are smaller and less representative of the population than the norming sample. Although not optimal, as long as the range of scores for $Test_1$ in the CRV study includes the value of the client's score, this information is better than having no CRV studies and making direct comparison of the scores.

Third, when encountering a situation where there are no published CRV studies that use the required tests or the studies do not provide all the required information, the situation becomes a bit trickier. If all that is provided are the tests' correlation, then score comparisons can be made with the realization that the predicted values will not be as precise as those that also used the test scores' means and SDs. Without at least knowing the correlation between the scores, we suggest that test scores should not be compared quantitatively. The reason for this is that without at least knowing the correlation between the tests scores, the way to compare the scores is the direct comparison method. Not only is this method more likely to produce overpredicted and underpredicted values, but this method provides no way of knowing the precision of the predicted score because it does not provide enough information to calculate CIs.

Fourth, psychologists should use their knowledge about a test's available CRV as part of the selection criteria they use when planning the tests they will administer for a given assessment. To find the CRV studies for a test, psychologists can examine the test's technical manual as well as other sources such as Google Scholar and PsycINFO. Although this extra work might appear to be a burden for practicing psychologists, it falls directly in line with other guidelines for ethical and evidence-based assessment practices (Hunsley & Mash, 2008; Joint Committee on Testing Practices, 2004). The European Federation of Psychology Associations provides a free

and relativity simple worksheet to use when making decisions about test use (Evers et al., 2013).

REFERENCES

Aiken, L. S., West, S. G., & Millsap, R. E. (2008). Doctoral training in statistics, measurement, and methodology in psychology: Replication and extension of Aiken, West, Sechrest, and Reno's (1990) survey of PhD programs in North America. *American Psychologist*, *63*, 32–50. doi:10.1037/0003-066X.63.1.32

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing* (3rd ed.). Washington, DC: Authors.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–560). Washington, DC: American Council on Education.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*, 307–310.

Bland, J. M., & Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, *8*, 135–160. doi:10.1177/096228029900800204

Castillo, J. M., Curtis, M. J., & Gelley, C. (2012). School psychology 2010—Part 2: School psychologists' professional practices and implications for the field. *NASP Communiqué*, *40*, 4–6.

Charter, R. A. (1996). Revisiting the standard errors of measurement, estimate, and prediction and their application to test scores. *Perceptual and Motor Skills*, *82*, 1139–1144. doi:10.2466/pms.1996.82.3c.1139

Childs, R. A., & Eyde, L. D. (2002). Assessment training in clinical psychology doctoral programs: What should we teach? What do we teach? *Journal of Personality Assessment*, *78*, 130–144. doi:10.1207/S15327752JPA7801_08

Clemence, A. J., & Handler, L. (2001). Psychological assessment on internship: A survey of training directors and their expectations for students. *Journal of Personality Assessment*, *76*, 8–47. doi:10.1207/S15327752JPA7601_2

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement*, *28*, 227–246. doi:10.1177/0146621604265031

Evers, A., Hagemeister, C., Høstmælingen, A., Lindley, P. A., Muñiz, J., & Sjöberg, A. (2013). *EFPA review model for the description and evaluation of psychological and educational tests* (Version 4.2.6). Brussels, Belgium: European Federation of Psychology Associations. Retrieved from http://www.efpa.eu/professional-development

Evers, A., Muñiz, J., Bartram, D., Boben, D., Egeland, J., Fernández-Hermida, J. R., … Urbánek, T. (2012). Testing practices in the 21st century. *European Psychologist*, *17*, 300–319. doi:10.1027/1016-9040/a000102

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flynn, J. R. (2012). *Are we getting smarter? Rising IQ in the twenty-first century*. New York, NY: Cambridge University Press.

Fouad, N. A., Grus, C. L., Hatcher, R. L., Kaslow, N. J., Hutchings, P. S., Madson, M. B., … Crossman, R. E. (2009). Competency benchmarks: A model for understanding and measuring competence in professional psychology across training levels. *Training and Education in Professional Psychology*, *3*(4), S5–S26. doi:10.1037/a0015832

Gresham, F. M. (2009). Interpretation of intelligence test scores in Atkins cases: Conceptual and psychometric issues. *Applied Neuropsychology*, *16*, 91–97. doi:10.1080/09084280902864329

Hall v. Florida, 572 U.S. (2014) Retrieved from http://www.supremecourt.gov/opinions/13pdf/12-10882_kkg1.pdf

Handler, L., & Smith, J. D. (2012). Education and training in psychological assessment. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of assessment psychology* (2nd ed., pp. 211–238). New York, NY: Wiley.

Haverkamp, B. E. (2013). Education and training in assessment for professional psychology: Engaging the 'reluctant student.' In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 2. Testing and assessment in clinical and counseling psychology* (pp. 63–82). Washington, DC: American Psychological Association.

Healy, M. J. R., & Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology*, *5*, 277–280. doi:10.1080/03014467800002891

Hunsley, J., & Mash, E. J. (2008). Developing criteria for evidence-based assessment: An introduction to assessments that work. In J. Hunsley & E. J. Mash (Eds.), *A guide to assessments that work* (p. 3–14). New York, NY: Oxford University Press.

Joint Committee on Testing Practices. (2004). *Code of fair testing practices in education*. Washington, DC: Author.

Lambert, M. J., & Vermeersch, D. A. (2013). Psychological assessment in treatment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 2. Testing and assessment in clinical and counseling psychology* (pp. 213–229). Washington, DC: American Psychological Association.

Neimeyer, G. J., Taylor, J. M., & Philip, D. (2010). Continuing education in psychology: Patterns of participation and perceived outcomes among mandated and nonmandated psychologists. *Professional Psychology: Research and Practice*, *41*, 435–441. doi:10.1037/a0021120

Norcross, J. C., & Karpiak, C. P. (2012). Clinical psychologists in the 2010s: 50 years of the APA Division of Clinical Psychology. *Clinical Psychology: Science and Practice*, *19*, 1–12. doi:10.1111/j.1468-2850.2012.01269.x

Ree, M. J., & Carretta, T. R. (2006). The role of measurement error in familiar statistics. *Organizational Research Methods*, *9*, 99–112. doi:10.1177/1094428105283192

Reynolds, C. R. (2008). Has any real understanding of measurement gone missing from the professional psychology curriculum? *The Score*, *30*(2), 3–4.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales*. Lutz, FL: Psychological Assessment Resources.

Reynolds, C. R., & Milam, D. A. (2011). Challenging intellectual testing results. In D. Faust (Ed.), *Coping with psychiatric and psychological testimony* (6th ed., pp. 311–334). New York, NY: Oxford University Press.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales* (5th ed.). Itasca, IL: Riverside.

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

Seashore, H. G. (1955). *Methods of expressing test scores* (Test Service Notebook 48). New York, NY: Psychological Corporation.

Shrout, P. E. (1998). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research*, *7*, 301–317. doi:10.1177/096228029800700306

Stedman, J. M., Hatch, J. P., & Schoenfeld, L. S. (2001). The current status of psychological assessment training in graduate and professional schools. *Journal of Personality Assessment*, *77*, 398–407. doi:10.1207/S15327752JPA7703_02

Swenson, E. V. (2013). Legal issues in clinical and counseling testing and assessment. In K. F. Geisinger, B. A. Bracken, J. F. Carlson, J. I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology: Vol. 2. Testing and assessment in clinical and counseling psychology* (pp. 83–99). Washington, DC: American Psychological Association.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest*, *1*, 1–26. doi:10.1111/1529-1006.001

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale: Technical and interpretive manual* (4th ed.). San Antonio, TX: Pearson.

Wechsler, D., Coalson, D. L., & Raiford, S. E. (2012). *Wechsler Preschool and Primary Scale of Intelligence technical and interpretive manual* (4th ed.). San Antonio, TX: The Psychological Corporation.

APPENDIX

Calculating Predicted Test Scores and Their Confidence Intervals

R Syntax[3]

 # CRV data
test1.mean <- 107.89 test1.sd <- 17.97 test2.mean <- 100.32 test2.sd <- 16.18 test.cor <- .76

 # test score
test1.score <- 100

 # predicted value
test2.predicted <- test2.mean +test.cor × (test2.sd/test1.sd) × (test1.score – test1.mean)

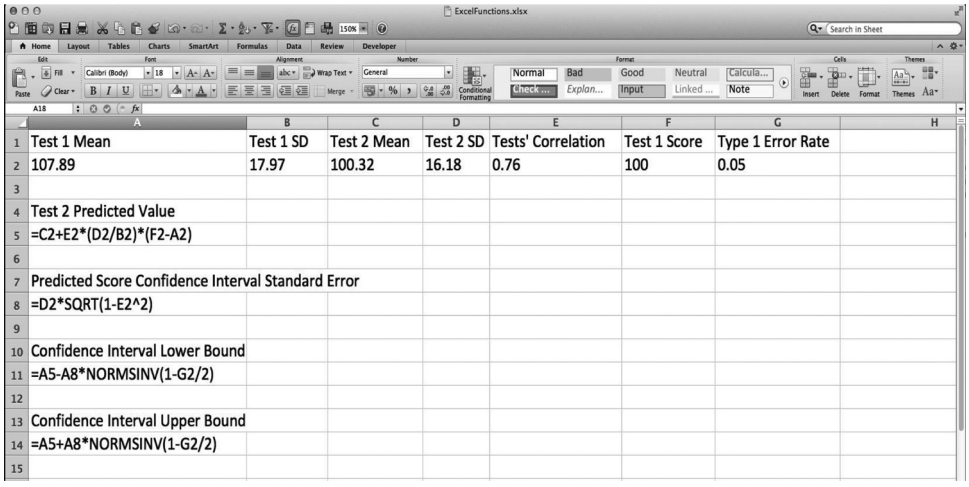
 # standard error
test2.predicted.se <- test2.sd × sqrt(1 – test.cor^2)

 # type one error rate
alpha <- .05

 # confidence interval lower bound
test2.predicted – test2.predicted.se × qnorm((1 – alpha/2))

 # confidence interval upper bound
test2.predicted +test2.predicted.se × qnorm((1 – alpha/2))

Microsoft Excel Syntax



[3]**R** is a free program that can be downloaded from http://www.r-project.org