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 & Karpia, 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.
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. The formula to transform raw scores to $Z$ scores is

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

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.

---

1This 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.
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

$$\text{Outcome variable} = \text{Intercept} + \text{Slope} \times \text{Predictor variable}$$

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 Test$_2$ hat. The predicted values for Test$_2$ are

<table>
<thead>
<tr>
<th>Test</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>RIAD</td>
<td>100.32</td>
<td>16.18</td>
</tr>
<tr>
<td>WAIS-III</td>
<td>102.90</td>
<td>14.90</td>
</tr>
<tr>
<td>WPPSI-IV</td>
<td>100.20</td>
<td>12.70</td>
</tr>
<tr>
<td>WAIS-III</td>
<td>107.00</td>
<td>18.70</td>
</tr>
<tr>
<td>WISC-III</td>
<td>107.89</td>
<td>17.97</td>
</tr>
<tr>
<td>WISC-IV</td>
<td>100.00</td>
<td>15.20</td>
</tr>
<tr>
<td>WPPSI-IV</td>
<td>100.20</td>
<td>12.70</td>
</tr>
<tr>
<td>WAIS-III</td>
<td>100.00</td>
<td>15.20</td>
</tr>
<tr>
<td>WAIS-IV</td>
<td>100.00</td>
<td>15.20</td>
</tr>
<tr>
<td>SB-V</td>
<td>101.50</td>
<td>14.40</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reynolds &amp; Kamphaus (2003)</td>
</tr>
<tr>
<td>Wechsler (2008)</td>
</tr>
<tr>
<td>Wechsler, Coalson, &amp; Raiford (2012)</td>
</tr>
<tr>
<td>Roid (2003)</td>
</tr>
</tbody>
</table>

calculated as

\[ \text{Test}_2 = a + b \times \text{Test}_1 \]  

(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:

\[ \text{Test}_2 \text{ variable} = a + b \times \text{Test}_1 \text{ predictor} + e \text{ error} \]  

(4)

Combining Equations 3 and 4 shows the error term is just the difference between the actual value of Test2 and the predicted value of Test2 (i.e., Test2 - Test2). Consequently, the best values to use for a and b are those that make the error term as small as possible, which then make Test2 and Test2 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 Test2 is expected (i.e., on average) to change when Test1 increases by 1 point. The intercept, a, is the expected value of Test2 when Test1 is 0. The predicted Test2 score, Test2, is still calculated using Equation 3, but its interpretation is now the expected value of Test2 for all the respondents who earned a specific score on Test1.

Relationship Between Regression Slope and Correlation

Regression and correlation are intimately related to each other. To understand this, imagine transforming the scores from Test1 and Test2 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.3

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:

\[ b = r_{12} \times \frac{SD_2}{SD_1} \]  

(5)

where \( r_{12} \) is the correlation between Test1 and Test2, and SD1 and SD2 are the SDs for Test1 and Test2, respectively (Cohen et al., 2003). To use Equation 5, the data for all three statistics (i.e., \( r_{12} \), SD1, and SD2) should all come from the same sample.

3In some texts, the slope estimated using Z scores is called a standardized regression coefficient.
PREDICTING TEST₂ SCORES FROM TEST₁ 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 Test₂ from a score on Test₁. To do so requires five values: (a) mean of Test₂ (M₂); (b) SD of Test₂ (SD₂); (c) mean of Test₁ (M₁); (d) SD of Test₁ (SD₁); and (e) the correlation between Test₂ and Test₁ (r₁₂). 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:

\[
\hat{\text{Test}_2} = \frac{M_2}{\text{intercept}} + \frac{r_{12} \times SD_2}{SD_1} \times (\text{Test}_1 - M_1)
\]

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 (Test₂): (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 (\(\sigma_{\text{Test}_2}\)). Combining those three parts produces a (1 - alpha)% CI:

\[
\hat{\text{Test}_2} \pm \sigma_{\text{Test}_2} \times N_{alpha/2}
\]

where \(N_{alpha/2}\) 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 \(\sigma_{\text{Test}_2}\) can be approximated by

\[
\sigma_{\text{Test}_2} \approx SD_2 \times \sqrt{1 - r_{12}^2}
\]

where \(SD_2\) is the SD of Test₂, and \(r_{12}^2\) is the squared correlation between Test₁ and Test₂ (Crocker & Algina, 1986). Notice in Equation 8 that the size of \(\sigma_{\text{Test}_2}\), 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 Test₂ 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 Test₂ 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 how the SDs are wrong. If the SD of Test₂ is less than the SD of Test₁, then the predicted values move away from their mean value. Thus, Test₂ values are underpredicted when Test₁ scores are below the mean and overpredicted when Test₁ scores are above the mean. If the SD of Test₂ is greater than the SD of Test₁, then the predicted values for Test₂ move close to the mean. Thus, Test₂ scores will be overpredicted when Test₁ scores are below the mean and underpredicted when Test₁ 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 Test₂ based on Test₁ scores. The mean and SD for Test₁ are 100 and 15, respectively, and the obtained scores on Test₁ are either 100, 70, or 130 (representing the mean and values 2 SDs above and below the mean). The SDs for Test₂ are either the same as Test₁ (i.e., SD = 15), lower than Test₁ (i.e., SD = 10), or higher than Test₁ (i.e., SD = 15). The correlations between Test₁ and Test₂ 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 Test₂ are relatively close to the Test₁ scores, although the predicted values using the .70 correlation are closer toward Test₁’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 Test₁ and Test₂ 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 Test₂ is 10
Comparing Test Scores

TABLE 2
Sample Predicted Values for Test2

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Test1</th>
<th>M1</th>
<th>SD1</th>
<th>Pred1</th>
<th>LB15</th>
<th>UB15</th>
<th>Off</th>
<th>Correlation = .90</th>
<th>Pred2</th>
<th>LB95</th>
<th>UB95</th>
<th>Off</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>70</td>
<td>100</td>
<td>15</td>
<td>73</td>
<td>60</td>
<td>86</td>
<td>0</td>
<td></td>
<td>79</td>
<td>58</td>
<td>100</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>100</td>
<td>15</td>
<td>100</td>
<td>87</td>
<td>113</td>
<td>0</td>
<td></td>
<td>100</td>
<td>79</td>
<td>121</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>130</td>
<td>100</td>
<td>15</td>
<td>127</td>
<td>114</td>
<td>140</td>
<td>0</td>
<td></td>
<td>121</td>
<td>100</td>
<td>142</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>70</td>
<td>90</td>
<td>15</td>
<td>63</td>
<td>50</td>
<td>76</td>
<td>-10</td>
<td></td>
<td>69</td>
<td>48</td>
<td>90</td>
<td>-10</td>
</tr>
<tr>
<td>5</td>
<td>100</td>
<td>90</td>
<td>15</td>
<td>90</td>
<td>77</td>
<td>103</td>
<td>-10</td>
<td></td>
<td>90</td>
<td>69</td>
<td>111</td>
<td>-10</td>
</tr>
<tr>
<td>6</td>
<td>130</td>
<td>90</td>
<td>15</td>
<td>117</td>
<td>104</td>
<td>130</td>
<td>-10</td>
<td></td>
<td>111</td>
<td>90</td>
<td>132</td>
<td>-10</td>
</tr>
<tr>
<td>7</td>
<td>70</td>
<td>110</td>
<td>15</td>
<td>83</td>
<td>70</td>
<td>96</td>
<td>10</td>
<td></td>
<td>89</td>
<td>68</td>
<td>110</td>
<td>10</td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>110</td>
<td>15</td>
<td>110</td>
<td>97</td>
<td>123</td>
<td>10</td>
<td></td>
<td>110</td>
<td>89</td>
<td>131</td>
<td>10</td>
</tr>
<tr>
<td>9</td>
<td>130</td>
<td>110</td>
<td>15</td>
<td>137</td>
<td>124</td>
<td>150</td>
<td>10</td>
<td></td>
<td>131</td>
<td>110</td>
<td>152</td>
<td>10</td>
</tr>
<tr>
<td>10</td>
<td>70</td>
<td>100</td>
<td>10</td>
<td>82</td>
<td>73</td>
<td>91</td>
<td>9</td>
<td></td>
<td>86</td>
<td>72</td>
<td>100</td>
<td>7</td>
</tr>
<tr>
<td>11</td>
<td>100</td>
<td>100</td>
<td>10</td>
<td>100</td>
<td>91</td>
<td>109</td>
<td>0</td>
<td></td>
<td>100</td>
<td>86</td>
<td>114</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>130</td>
<td>100</td>
<td>10</td>
<td>118</td>
<td>109</td>
<td>127</td>
<td>-9</td>
<td></td>
<td>114</td>
<td>100</td>
<td>128</td>
<td>-7</td>
</tr>
<tr>
<td>13</td>
<td>70</td>
<td>90</td>
<td>10</td>
<td>72</td>
<td>63</td>
<td>81</td>
<td>-1</td>
<td></td>
<td>76</td>
<td>62</td>
<td>90</td>
<td>-3</td>
</tr>
<tr>
<td>14</td>
<td>100</td>
<td>90</td>
<td>10</td>
<td>90</td>
<td>81</td>
<td>99</td>
<td>-10</td>
<td></td>
<td>90</td>
<td>76</td>
<td>104</td>
<td>-10</td>
</tr>
<tr>
<td>15</td>
<td>130</td>
<td>90</td>
<td>10</td>
<td>108</td>
<td>99</td>
<td>117</td>
<td>-19</td>
<td></td>
<td>104</td>
<td>90</td>
<td>118</td>
<td>-17</td>
</tr>
<tr>
<td>16</td>
<td>70</td>
<td>110</td>
<td>10</td>
<td>92</td>
<td>83</td>
<td>101</td>
<td>19</td>
<td></td>
<td>96</td>
<td>82</td>
<td>110</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>100</td>
<td>110</td>
<td>10</td>
<td>110</td>
<td>101</td>
<td>119</td>
<td>10</td>
<td></td>
<td>110</td>
<td>96</td>
<td>124</td>
<td>10</td>
</tr>
<tr>
<td>18</td>
<td>130</td>
<td>110</td>
<td>10</td>
<td>128</td>
<td>119</td>
<td>137</td>
<td>1</td>
<td></td>
<td>124</td>
<td>110</td>
<td>138</td>
<td>3</td>
</tr>
<tr>
<td>19</td>
<td>70</td>
<td>100</td>
<td>20</td>
<td>64</td>
<td>47</td>
<td>81</td>
<td>-9</td>
<td></td>
<td>72</td>
<td>44</td>
<td>100</td>
<td>-7</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>100</td>
<td>20</td>
<td>100</td>
<td>83</td>
<td>117</td>
<td>0</td>
<td></td>
<td>100</td>
<td>72</td>
<td>128</td>
<td>0</td>
</tr>
<tr>
<td>21</td>
<td>130</td>
<td>100</td>
<td>20</td>
<td>136</td>
<td>119</td>
<td>153</td>
<td>9</td>
<td></td>
<td>128</td>
<td>100</td>
<td>156</td>
<td>7</td>
</tr>
<tr>
<td>22</td>
<td>70</td>
<td>90</td>
<td>20</td>
<td>54</td>
<td>37</td>
<td>71</td>
<td>-19</td>
<td></td>
<td>62</td>
<td>34</td>
<td>90</td>
<td>-17</td>
</tr>
<tr>
<td>23</td>
<td>100</td>
<td>90</td>
<td>20</td>
<td>90</td>
<td>73</td>
<td>107</td>
<td>-10</td>
<td></td>
<td>90</td>
<td>62</td>
<td>118</td>
<td>-10</td>
</tr>
<tr>
<td>24</td>
<td>130</td>
<td>90</td>
<td>20</td>
<td>126</td>
<td>109</td>
<td>143</td>
<td>-1</td>
<td></td>
<td>118</td>
<td>90</td>
<td>146</td>
<td>-3</td>
</tr>
<tr>
<td>25</td>
<td>70</td>
<td>110</td>
<td>20</td>
<td>74</td>
<td>57</td>
<td>91</td>
<td>1</td>
<td></td>
<td>82</td>
<td>54</td>
<td>110</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>100</td>
<td>110</td>
<td>20</td>
<td>110</td>
<td>93</td>
<td>127</td>
<td>10</td>
<td></td>
<td>110</td>
<td>82</td>
<td>138</td>
<td>10</td>
</tr>
<tr>
<td>27</td>
<td>130</td>
<td>110</td>
<td>20</td>
<td>146</td>
<td>129</td>
<td>163</td>
<td>19</td>
<td></td>
<td>138</td>
<td>110</td>
<td>166</td>
<td>17</td>
</tr>
</tbody>
</table>

Note. The mean and standard deviation (SD) of Test1 are 100 and 15, respectively. M2 = mean of Test2; SD2 = SD of Test2; Pred2 = predicted score on Test2 given Test1 score; LB95 = lower bound of 95% confidence interval (CI); UB95 = 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 Test1. Consequently, the predicted Test2 scores are 10 points lower than those from their baseline scores. In Comparisons 7–9, the mean for Test2 is 10 points greater than that of Test1, making the predicted scores on Test2 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 Test2 is one third smaller than that of Test1 (i.e., the SD is 10 instead of 15). For Comparisons 10 through 12, there are no mean differences. For an average score on Test1 (Comparison 11), the predicted value for Test2 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 Test1 deviates from its mean, however, the predicted values for Test2 begin to move toward their mean. In Comparison 10, the score for Test1 is 70. With a correlation of .90, the predicted value for Test2 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 Test2 that is one third smaller than that for Test1, but also add that the mean for Test2 is 10 points lower than that of Test1. For an average score on Test1 (i.e., Comparison 14), the predicted value of Test2 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 Test1 deviates from its mean, however, the predicted values for Test2 move toward their means. For example, Comparison 13 has a score of 70 on Test1. 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 Test1 is 130. With a correlation of .90, the predicted value for Test2 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 Test1 produce larger differences from the baseline than do high scores.

Integrating the results from Comparisons 13 through 18 shows that when the values of Test1 are close to the actual mean of Test2, and the actual SD of Test2 is smaller than that of Test1, 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 Test1 depart from the actual mean of Test2, then the amount of the predicted values for Test2 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 Test2 is one third larger than that for Test1. Consequently, the results in Comparisons 22 through 27 show the opposite effect of Comparisons 13 through 18. When the value of Test1 is far from the actual mean of Test2, 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 Test1 get closer to the actual mean of Test2, then the amount of the predicted values for Test2 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:

\[
\sigma_{\text{RIAS}} \approx 16.18 \times \sqrt{1 - .76^2} = 10.52
\]

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:

\[
\sigma_{\text{WISC-III}} \approx 17.97 \times \sqrt{1 - .76^2} = 11.68
\]

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 Test1 based on Test2 will be exactly the same as the predicted values for Test2 based on Test1.
The predicted RIAS CIX 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.

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.
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


Calculating Predicted Test Scores and Their Confidence Intervals

**R Syntax**

```r
# 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**

1. Test 1 Mean  107.89
2. Test 1 SD  17.97
3. Test 2 Mean  100.32
4. Test 2 SD  16.18
5. Tests' Correlation  0.76
6. Test 1 Score  100
7. Type 1 Error Rate  0.05

---

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