Informed Test
Component Weighting

Lawrence M. Rudner

Maryland Assessment Research
Center for Education Success

A paper commissioned by the Maryland State Department of Education.

March 20, 2000

Department of Measurement, Statistics and Evaluation
University of Maryland, College Park 20742
Informed Test Component Weighting

Lawrence M. Rudner

University of Maryland, College Park

How can we combine a multiple-choice assessment with a performance assessment to yield a single score? Should we give more weight to a longer, but less reliable, performance assessment? Or should we add less weight?

In many assessment situations, multiple tests or subtests are administered and the results are combined to form a single composite score. For example, today’s commercially available achievement tests report a composite reading score formed by combining literal comprehension with inferential comprehension components. Other common uses of composite scoring can be found in the SAT in which the verbal and mathematical reasoning ability components are combined to form a composite score, in interest inventories and personality tests, which often combine tests of different traits and in employment decision-making, which is generally based on a combination of factors. Today, many large-scale assessment programs are combining components that incorporate traditional multiple-choice items with components that incorporate newer performance items.

The manner in which composite scores are formed raises a variety of methodological and policy issues. For example, do we add more weight to a longer, but less reliable, performance assessment? Or should we add less weight? This problem is not new and a host of methods have been proposed. A half of a century ago, Gulliksen (1950) devoted a 50 page chapter to the topic. Thirty years ago, Wang and Stanley (1970) wrote a comprehensive 40 page review. Dawes (1976) revised the problem in his classic article on equally weighted measures. More recently Wainer and Thissen (1993) discussed the issue in the context of combining multiple-choice and constructed-response tests. Focusing on the reliability of composite scores, most of the literature suggests that weighting doesn’t matter. Weighting does not appear to reduce overall error and the results of different weighting methods are often highly correlated. The effect of weights on the validity of the composite score, however, has not been adequately addressed.

This paper identifies and logically evaluates alternative methods for weighting tests. Formulas are presented for composite reliability and validity as a function of component weights. Concluding that weighting can make a big difference when combining a highly reliable test, such as a lengthy multiple-choice test, with a less reliable test, such as a short constructed-response test, this paper suggests a rational process that identifies and considers trade-offs in determining weights.

Component Weighting Methods

As Gulliksen (1950) points out, and Wainer and Thissen (1993) underscore:

It should be noted that it is not possible to dodge the weighting problem if any decisions are to be made. Occasionally, we hear the suggestion that scores simply be added
together without bothering about the problem of weighting. No matter what scores we add, the problem is not avoided (p 312).

Weighting methods are either implicit or explicit. This section identifies and briefly evaluates some of the more theoretically and conceptually appealing approaches for determining weights. The interested reader is referred to Wang and Stanley (1970) and Gulliksen (1950) for more thorough reviews of non-IRT methods.

Implicit Approaches

Adding Raw Scores - Perhaps the simplest approach to combining test components is to add together the total number of correct responses. In theory, if the tests are developed following a blueprint, the number of items within each domain should be a fair representation of the domain’s relative importance. Adding raw scores, however, fails to recognize differences in variation across domains and across items. By adding raw scores, the implicit weights are not the relative importance of each domain, but rather the variance of domain scores. For this reason, many introductory measurement textbooks advise against combining raw scores and instead recommend the use of standard scores. Another issue is that equal item weighting fails to consider differences in item importance. A lengthy algebra solution cannot be considered equal to recognizing an inequality on a multiple-choice test.

IRT Modeling - Rather than tackling the issue, one could simultaneously calibrate the items across all components and use an IRT model to estimate each examinee ability on the composite scale. The implicit weights vary according to the IRT model. Deriving theta from simultaneously calibrated one-parameter IRT items is equivalent to summing the item scores. Deriving theta from a two-parameter model is equivalent to weighting by the discrimination of the items within each component. The three-parameter model would be influenced by both the discrimination parameter and the theta estimate (Lord, 1980 p. 74-77). With all IRT models, the implicit weighting optimizes the fit of the model to the data.

Explicit Approaches

Weighting by Difficulty - Instructors often weight items or sections of classroom tests based on their feel for the task difficulty. The same concept can be applied with empirical data. This approach appears to be attractive because it provides additional reward for mastering particularly difficult concepts. However, the converse is also true: this method punishes students more severely for missing more difficult items. Weighting by easiness just reverses the penalties.

Reliability Weighting - Giving more reliable components heavier weights is intuitively appealing. The error associated with the composite score would be less if the more reliable measure were more heavily weighted. As seen in equation (1) below, however, simply weighting by component reliability will not maximize composite reliability. Weights for maximizing composite score reliability can be determined using Monte-Carlo techniques and equation (1) or by setting the first derivative of (1) with respect to \( w_1/w_2 \) to zero and solving. Wang (1998) also presents a workable
solution based on the variance-covariance matrix. We will argue later, however, that maximizing reliability is not necessarily a worthwhile goal.

**Validity Weighting** - A variety of methods can be used to maximize the validity of the composite scores. Multiple regression provides component weights that maximize the correlation of the composite with an external criterion. There is the well-documented issue of shrinkage. The regression optimizes weights for the given data set. The weights may be less than optimal for other datasets. The validity coefficients themselves could be used as weights. This, however, would be even less optimal because it fails to consider the intercorrelations among the predictors. Further, determining a criterion in order to estimate validity coefficients is not always straightforward. As with maximizing reliability, maximizing validity may not be the most desirable goal. Further the determination of validity presents notable challenges.

**Formulas for Composite Reliability and Validity**

Let random variables $X = [X_1, X_2]$ denote two components and random variable $Y$ denote scores on a criterion variable. Further let $L = w_1X_1 + w_2X_2$ denote the weighted composite test score. To simplify calculations, set variances equal to unity. As a result $\hat{\sigma}_1 = \hat{\sigma}_2 = \hat{\sigma}_y = 1.0$, $\hat{\rho}_{12}$ and $\hat{\rho}_{Yi} = \hat{\rho}_{Yi}$ for $i=1,2$.

Wang and Stanley (1970, p. 672) provide a general formula for the reliability of a composite $L$ composed of $n$ variables. Solving for $n=2$ variables and simplifying, we have the reliability of a composite as a function of the weights, component reliabilities and the positive correlation between the components.

$$\rho_{LL'} = \frac{w_1^2\rho_{11} + w_2^2\rho_{22} + 2w_1w_2\rho_{12}}{w_1^2 + w_2^2 + 2w_1w_2\rho_{12}} \quad (1)$$

From (1), we can derive that:

- a) The lowest possible value for the composite reliability is the reliability of the less reliable component.
- b) If the components are correlated, then the composite reliability can be higher than the reliability of either component.
- c) If the component reliabilities are the same, then the composite reliability is maximum when the weights are the same.

Winer (1971, p. 105) provides an equation for the squared product moment correlation between criterion variable $Y$ and $L$ as the product of the variance-covariance matrix, the scalar array of weights, and the scalar array of component-criterion variable covariances. Solving for $n=2$ predictor variables and simplifying yields the multiple correlation of a composite with a criterion
variance as a function of the weights, component validities and the correlation between the components.

\[ \rho_{yl} = \frac{w_1 \rho_{y1} + w_2 \rho_{y2}}{\sqrt{w_1^2 + w_2^2 + 2w_1w_2\rho_{12}}} \]  \hspace{1cm} (2)

Linear regression provides the weights that maximize the correlation between a composite and a criterion. Thus, the square root of the Multiple R associated with linear regression provides the maximum value for (2).

\[ \max \rho_{yl} = \sqrt{\frac{\rho_{y1}^2 + \rho_{y2}^2 - 2\rho_{y1}\rho_{y2}\rho_{12}}{1 - \rho_{12}^2}} \]  \hspace{1cm} (3)

From (2) and (3), we can deduce:

a) The lowest possible value for the composite validity is the validity of the less valid component.

b) The composite validity can be higher than the validity of either component.

c) If the component validities are the same then the composite validity is maximum when the weights are the same.

d) The maximum possible composite validity increases as the component correlation decreases.

**Example**

This section provides an example to illustrate the effect of component weights on the composite test reliability and composite test validity. The analysis is based on the Biology AP examination as reported by Wainer and Thissen (1993). Component 1 is comprised of multiple-choice items with a reliability of .93. Component 2 is comprised of constructed-response items with a reliability of .68. The correlation of the two is .73 (0.92 unattenuated).

Table 1 shows there is very little change in the reliability as the ratio of the weights change from \( \infty \) to 1:1. The reliability starts to drop precipitously as extra weight is given to the less reliable constructed-response component. In this example, the inclusion of the constructed-response component hurts reliability.
Table 1

<table>
<thead>
<tr>
<th>weight $W_1/W_2$</th>
<th>$\infty$</th>
<th>8/1</th>
<th>4/1</th>
<th>2/1</th>
<th>1/1</th>
<th>1/2</th>
<th>1/4</th>
<th>1/8</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>reliability $\tilde{R}_{LL}$</td>
<td>.93</td>
<td>.93</td>
<td>.94</td>
<td>.92</td>
<td>.89</td>
<td>.83</td>
<td>.77</td>
<td>.73</td>
<td>.68</td>
</tr>
</tbody>
</table>

Looking at reliability data from a large number of composite tests where the longer multiple-choice component had a higher reliability, Wainer and Thissen (1993) concluded that “whatever is being measured by the constructed-response section is measured better by the multiple choice section. These seven tests are but a sample. We have never found any test ... for which this is not true.” These powerful words will hold as long as both sections have comparable validity and the constructed response section has lower reliability.

But what if we value the content of the constructed-response more? That is, what if that section is more highly correlated with a criterion than the multiple-choice section? McCornack (1956) discussed this more than 40 years ago. The studies of his day mostly looked at the effect of weighting components with high reliability on the composite reliability and concluded that weighting doesn’t matter. McCornack criticized these studies for not considering composite validity. The criticism also applies to Wainer and Thissen.

Differential component validity for the Biology AP examination is modeled in Table 2 and Figure 1. Suppose we have a highly valid and reliable criterion, the correlation of the more reliable multiple-choice component with the criterion is 0.60 (.62 unattenuated), and the validity coefficient for the less reliable constructed-response component is .80 (.97 unattenuated). Figure 1 plots formulas (1) and (2) for different weight ratios and the values shown in Table 2. Variable $w_1$ indicates the weight for the multiple-choice component and $w_2$ indicates the weight for the constructed-response component. A logarithmic scale is used for the horizontal axis as the ratios less than 1 are the inverse of the ratios greater than 1. Figure 1 shows that composite reliability rises steadily as more weight is given to the more reliable multiple-choice component and starts to level off when $w_1=1.5w_2$. Composite validity slowly drops as $w_1$ approaches .7$w_2$ and then starts to drop more precipitously. Clearly, in this example, the weights have a profound effect on composite validity and composite reliability.
Contrary to the oft-quoted principle that square root of the reliability places an upper limit on validity, composite validity is increasing while composite reliability is decreasing in this example. This is not a mistake in formulas (1) and (2). That principle doesn’t apply here because the components are not highly correlated. As astutely noted by Feldt (1997), the principle applies when the reduced reliability represents a similar, but shorter test. Here, by changing the weights, we have changed the essential character of the test: our test now measures something different -- the true scores represent a different construct.

**Discussion**

We have argued that implicit weighting methods may not yield the desired results and that explicit weighting can seriously impact composite validity and composite reliability. Further, weighting...
can have unsatisfactory consequences. Maximizing reliability can lead to lower validity. Maximizing validity can lead to lower reliability.

The question remains, how does one weight two components? As argued by Kennedy and Walstad (1997), weighting should be a rational process evaluating contributions and the trade-offs. In Figure 1, for example, if one feels consistency is extremely important and that a validity coefficient of .75 is adequate, then a \( \frac{w_1}{w_2} \) between 1.2 and 2.0 should be supported. Conversely, if one feels validity is more important and that a reliability of about .75 is adequate, then weighting component 2 more heavily with a \( \frac{w_1}{w_2} \) of about .5 should be adopted. In both cases, the trade-offs between reliability and validity can be rationally considered.

It should be noted from formulas (1) and (2) that the ends of the composite validity and composite reliability curves asymptotically approach the individual component validities and reliabilities. As the correlation between the components goes up, the curves become less peaked. As the correlation approaches unity, one could simply maximize reliability. Thus, if the component validities are each satisfactory or the components are fairly intercorrelated, the different weights will not make much difference on composite validity. We suspect this will be the case with many large scale assessments that incorporate alternative item types. Nevertheless, one needs estimates of component validity, or a very high component correlation, before one can dismiss the effect of component validity on the validity of the composite scores.

The absence of a natural criterion variable does not mean data cannot be collected and used to help inform the decision. Surrogate markers for clinically defined end-points (Prentice, 1989) are commonly used in medical research. Traditional approaches for conducting content validity studies based on ratings of item-objective congruence and relevance are applicable. A variety of new quantitative techniques for conducting content validity studies based on multidimensional scaling have been offered (Sireci, 1998). Value judgements can also be employed instead of the criterion measure, as long as the value judgement is a statement of worth and not perceived difficulty. Again, formulas (1) and (2) and a graph similar to Figure 1 can be used to help make the weighting decision policy.

References


Lawrence Rudner is the Director of the ERIC Clearinghouse on Assessment and Evaluation and the Assistant Director of the Maryland Assessment Research Center for Education Success at the University of Maryland, Department of Measurement, Statistics, and Evaluation, 1129 Shriver Lab, College Park, MD 20742. His specializations are measurement theory and information science.

This research was sponsored by the Maryland State Department of Education through a contract with the Maryland Assessment Research Center for Education Success, College of Education, University of Maryland, College Park. The opinions expressed do not necessarily reflect the positions of the Maryland State Department of Education and no official endorsement should be inferred.