Internal Consistency: Do We Really Know What It Is and How to Assess it?

Wei Tang, Ying Cui, CRAME, Department of Educational Psychology
University of Alberta, CANADA

The term “internal consistency” has been widely used but controversially defined. Cronbach (1951) used internal consistency and homogeneity interchangeably and claimed: “an internally consistent or homogeneous test should be independent of test length”. However, many other researchers (Green, Lissitz & Mulaik, 1977; McDonald, 1981; Miller, 1995; Schmitt, 1996) used the term internal consistency to refer to the interrelatedness of items, and distinguished internal consistency from homogeneity by claiming that homogeneity refers to the unidimensionality of a set of items. Another popular use of internal consistency is related to the reliability estimate of a test based on a single administration, which is traditionally called internal consistency reliability (Haertel, 2006). In this sense, internal consistency has been often used to denote a group of methods that are intended to estimate the reliability of a single-administrated test (Hattie, 1995).

As illustrated above, the term “internal consistency” is associated with different meanings. Just as Sijtsma (2009) concluded, “Internal consistency has not been defined that explicitly, far from it”. Because of the vague and implicit definition of this term, the chaos in interpreting and using internal consistency has existed for a long time. The confusion in the concept of internal consistency may cause problems in understanding, measuring and applying other related psychometric properties of a test. Besides, it hampers the development of new and better indices for measuring test internal consistency. Therefore, an explicit and complete definition of the term
“internal consistency” is badly needed. To better understand and define this term, different definitions or interpretations of the term “internal consistency” and the corresponding measures for assessing internal consistency are reviewed and discussed in the following sections. Finally, an explicit definition of test internal consistency is provided and appropriate measures are recommended.

**What Is Internal Consistency?**

**Internal Consistency = Internal Consistency Reliability?**

Internal consistency has long been used as the synonym of internal consistency reliability and thus the indices for measuring internal consistency reliability, such as Kuder and Richardson (1937)’s KR–20 and coefficient alpha (Cronbach, 1951), have become widely used measures of internal consistency. Interpretation of alpha as a measure of internal consistency has gained more foothold in practical test construction and test use (Sijtsma, 2009). However, some researchers (Cortina, 1993; Green et al, 1977; Nunnally, 1978) have pointed out the inappropriateness of using internal consistency reliability coefficients to measure internal consistency. For example, Nunnally wrote: “Estimates of reliability based on the average correlation among items within a test are said to concern the ‘internal consistency.’ This is partly a misnomer, because the size of the reliability coefficient is based on both the average correlation among items (the internal consistency) and the number of items. Coefficient alpha is the basic formula for determining the reliability based on internal consistency.”

To better distinguish internal consistency and internal consistency reliability, we may first review the definition of reliability. The reliability of measurement results was initially defined by Spearman (1904) as the ratio of true score variance to
observed score variance in classical test theory. Reliability is the quantification of the consistency of results from a measurement procedure across replications. A test or measure is considered perfectly reliable if we get the same results repeatedly. The reliability of measurement results can be estimated by calculating the correlation between the total scores on two independent administrations or two parallel forms of the test. Although the definition of reliability requires data from repeated testing, such data are rarely available in practice. Thus, reliability coefficient is normally estimated based on scores from a single test administration, which is referred to as internal consistency reliability by Cronbach (1951). Intuitively, internal consistency reliability concerns the consistency of behavior within a very limited time interval, i.e., the time interval during which the items in the test are being responded to (Horst, 1953). And this “short term” reliability is comparable to the correlation between two tests taken within a short time interval.

Internal consistency in the term “internal consistency reliability” does not work as the synonym of reliability but the qualifier of the reliability, which implies the reliability is estimated under the condition of internal consistency, that is, when the test is internally consistent, the reliability of the test can be estimated by these reliability coefficients. An internal consistent test can be interpreted as a test with all items measure the same underlying construct in the same way. In other words, test items are all parallel in an internally consistent test. Internal consistency reliability such as KR–20 and coefficient alpha are not designed to measure internal consistency but the reliability under the condition of the internal consistency. Using these reliability estimates for measuring internal consistency is inappropriate.

Therefore, although both reliability and internal consistency are determined by the degree of interitem correlation, these two terms are distinctive in nature.
Reliability refers to the consistency of a whole test across replications whereas internal consistency refers to the consistency among the responses to the items within a test. A test that is not internally consistent could be reliable, and vice versa. The coefficients for internal consistency reliability should not be used directly as indices for internal consistency also because reliability estimates are all functions of item number. Specifically, reliability increases as item number increases. Different from reliability, internal consistency of a test should be independent of its length (Cronbach, 1951). For example, a test with 10 items may have the same internal consistency as it does when increased to 20 items, even though the reliability of the test is higher then. If test developers use many items to measure a single construct without increasing test internal consistency, this can serve only to increase reliability at the expense of precision and parsimony.

**Internal Consistency = Average Interitem Correlations?**

Conceptually, internal consistency measures whether items on the same test (or the same subscale on a larger test) that intend to measure the same construct produce consistent scores. If two items are designed to measure the same construct, a testee should answer both questions in the same way, which would indicate that the test has internal consistency. Thus, some researchers (Cortina, 1993) defined internal consistency as a measure based on the degree of correlations between different items on the same test (or the same subscale on a larger test). Since the correlations between items are, most often than not, different, using the average interitem correlation seems a simple and direct approach to represent the degree of correlations between different items. Cronbach (1951), to the authors’ knowledge, is the first one who proposed to use mean interitem correlation to measure internal consistency. He derived an index
(Cronbach’s $\bar{r}_{ij}$) specifically for assessing internal consistency, though no explicit
definition of internal consistency is provided. Cronbach’s $\bar{r}_{ij}$ (1951) is derived by
applying the Spearman-Brown formula to alpha, thereby estimating the mean
correlation between items. Since then, internal consistency has been interpreted and
measured by the mean interitem correlation by some researchers and practitioners
(e.g., Briggs & Cheek, 1986; Nunnally, 1978).

However, we should be aware that the “consistent scores” does not necessarily
mean the scores are same or similar across items. For example, in a psychological test,
if all respondents expressed agreement with the statements "I like to study with a
partner" and disagreement with the statement "I hate to study with a partner", this is
still an indicative of perfect internal consistency of the test. Although the two items
are negatively correlated, the degree of the correlation is as high as 1. Suppose there
are more items like that in the test, then the mean interitem correlations could be close
to zero because the positive interitem correlations cancel out the negative ones. To
overcome that problem, the practitioners can either recode the raw scores in the
opposite order or use the mean of the absolute value of the correlations between items
as the indicator for internal consistency.

One disadvantage of mean interitem correlation is that this measure is influenced
by the extreme values or outliers. In addition, it does not reflect the variability among
the interitem correlations. Therefore, the use of mean interitem correlation will be
problematic when the interitem correlations follow a skewed distribution, especially
with some extreme outliers.

**Internal Consistency = General Factor Saturation?**

An alternative way of interpreting internal consistency is that it is the extent to
which all of the items of a test measure the same construct, that is, the general factor saturation (Revelle, 1979) or the closeness to the unidimensionality (Sijtsma, 2009). The advantage of this perspective over the notion of the average degree of correlation among the items of a test is that the general factor saturation is not affected by the skewness of the distribution of item correlations or extreme values of item correlations. Besides, if the items of a test measure several unrelated latent traits, the average item correlation in such cases will be greater than zero whereas some subtests or subscales which measure the unrelated latent traits have the correlation of zero. Thus, using the average degree of item correlation to measure internal consistency is not an appropriate choice when a test is not unidimensional.

Whereas the ideal of measurement is for all items of a test to measure the same latent variable, it is hard to achieve for a single test in real practice. For example, IELTS (International English Language Testing System) has four subscales which measure four distinctive features of English: listening, reading, writing and speaking. Although the whole test is designed to measure the proficiency level of English as a second language (the general trait/construct), each module taps a lower order trait (the group trait/construct). Tests or measures with items reflecting a single factor in the higher order while items reflecting a variety of aspects mark a number of factors in the lower order (one factor for each aspect) is not uncommon in psychological or educational measurement, and hence the assumption of unidimensionality of a test is often unsatisfied. Nevertheless, if the subscales are highly correlated, it is factorially homogeneous (Revelle, 1979). Since high correlations among the subscales indicate the high order factor (the general factor) can be extracted and be dominant, test internal consistency is interpreted as the general factor saturation or the closeness to the unidimensionality. That also explains why some researchers (Cronbach, 1951;
Revelle, 1979) use homogeneity and internal consistency as synonyms.

As to the measures of test internal consistency defined in this way, different recommendations have been raised so far. Revelle (1979) recommended beta, the worst split-half reliability estimate, and pointed out: “in the case of a lumpy test (one with several large group factors), alpha overestimates the general factor saturation of the test and underestimates the total common factor saturation. Beta, on the other hand, gives a more appropriate estimate of the general factor saturation but severely underestimates the common factor saturation.” However, to find this worst split half requires trying all the possible splits, which is impractical when the test is just normal size, say, of 20 items. Since analytic method does not work, a heuristic is needed to estimate beta. ICLUST (Revelle, 1977; 1979), the program for hierarchical cluster analysis, was put forward to estimate beta. Sijtsma (2009) suggested using MRFA (minimum rank factor analysis, Ten Berge & Kiers, 1991) for assessing closeness of the covariance/correlation matrix to unidimensionality. By using MRFA, closeness of the 1-factor solution to unidimensionality is assessed by the ratio of the first eigenvalue of the estimated true score covariance matrix to the sum of all J (J is the number of variables) eigenvalues (Ten Berge & Sočan, 2004). Sijtsma used the percentage of the above ratio, expressed as the ECV (explained common variance), to represent the closeness to unidimensionality. Revelle and Zinbarg (2009) suggested the hierarchical coefficient omega (ωₜ) (McDonald, 1999; Zinbarg, Revelle, & Yovel, 2005) might be a more appropriate index of estimating internal consistency since ωₜ could indicate the extent to which all of the items in a test measure the general factor. ωₜ equals the ratio of the general factor variance to the total variance of the test, and thus a natural choice of measuring general factor saturation.
An Explicit Definition of Internal Consistency

Based on the above review and discussion of the term “internal consistency” and the corresponding measures for assessing internal consistency, internal consistency can be defined as a psychometric property of a test that describes whether the items in the test that are proposed to measure the same construct produce consistent scores, which is associated with the degree of general factor saturation and the degree of interitem correlations, but irrelevant to test length. If defined more stringently, internal consistency means all the items are parallel in a test or subtest. It can be shown that internal consistency has a stronger connection with validity than reliability if we look at the formulas that depict the relationship between validity and reliability (Judd, Smith & Kidder, 1991; Streiner, 2003) below:

\[
\text{Reliability} = \frac{\sigma_{CI}^2 + \sigma_{SE}^2}{\sigma_T^2}, \quad (1)
\]

and

\[
\text{Validity} = \frac{\sigma_{CI}^2}{\sigma_T^2}. \quad (2)
\]

The CI represents the construct of interest, the SE the systematic error. Therefore, if internal consistency is simply defined as the general factor saturation, measuring internal consistency is no difference from measuring construct validity. However, as stated above, internal consistency is a more complex concept than general factor saturation because it reveals more information about the internal structure of a test.

Comparison of the Indices for Measuring Internal Consistency

Method

To find out which of the recommended indices are appropriate measures of internal consistency, these indices are compared by using hypothetical data sets. The
main advantage of using artificial data is that the internal structures of hypothetical tests are known. Besides, tests with various internal structures can be analyzed and compared simultaneously, and thus the pattern or trend of the indices and the factors influencing the indices can be detected. The conditions are manipulated to resemble real psychological tests to overcome the disadvantage of hypothetical data: being artificial. Although psychological tests are diverse in terms of the content, such as achievement tests, aptitude tests, intelligence tests, personality tests, and interest inventories, they have several common points which can be utilized for data simulation. First, psychological tests often measure more latent traits or attributes instead of a single trait that is assumed in the classical test theory. For example, the questions on PDS (Personal Data Sheet) cover the topics such as excessive anxiety, depression, abnormal fears, impulse problems, etc (Woodworth, 1920). Second, the traits to be measured in a psychological test tend to be interrelated because they are the components of a more general construct. Considering these common characteristics in psychological tests, the model chosen for simulating real psychological tests is the bifactor model (Chen, West, & Sousa, 2006; Rindskopf & Rose, 1988; Schmid & Leiman, 1957; Yung, Thissen, & Mcleod, 1999), which includes not only a general factor F, but also K group factor(s),G_k, expressed as:

\[ X_i = \lambda_i F + \sum_{k=1}^{K} \alpha_{ik} G_k + E_i \]

Here, \( \lambda_i \) is the factor loading for the general factor F and \( \alpha_{ik} \) is the factor loading for the kth group factor \( G_k \) on the ith item; \( E_i \) is the residual score of the ith item.

To serve the purpose of comparison, the unifactor model is also included (i.e, \( \alpha_{ik} \) is set as zero). Totally, test data sets with five different internal structures are generated. 1) unifactor data with equal \( \lambda_i \); 2) unifactor data with unequal \( \lambda_i \); 3) bifactor data with larger \( \lambda_i \) than \( \alpha_{ik} \); 4) bifactor data with equal \( \lambda_i \) and \( \alpha_{ik} \). 5)
bifactor data with smaller $\lambda_i$ than $\alpha_{ik}$. To generate a simple bifactor structure, only two group factors are considered with equal group factor loadings, and the general factor loadings are also the same in bifactor data sets. Each structure is manipulated with three levels of items number (10, 20 and 40), and two levels of average interitem correlation (0.3 and 0.6) for the unifactor data and two subscales of bifactor data.

**Results**

**Unifactor Data**

Cronbach’s $\bar{r}_{ij}$ is exactly the same as the mean of interitem correlations. In fact, this index is equal to the average interitem correlation no matter whether the item covariance matrix is identical to the item correlation matrix as in the cases of this study. As shown in Table 1, Cronbach’s $\bar{r}_{ij}$ (shortened as C’s $\bar{r}_{ij}$) is only affected by the degree of factor loadings of the item, that is, the degree of interitem correlations. If the average interitem correlation does not change, the coefficient $\bar{r}_{ij}$ keeps constant as the item number increases from 10 to 40, regardless of the change in the discrepancy between interitem correlations. SD is the standard deviation of the correlations between items. When item factor loadings are the same, SD is zero. Under such condition, test length has no effect on SD. Under the conditions with heterogeneous loadings, test length has negligible effect on SD. As item number increases from 10 to 20, SD increases by 0.01.

Whereas alpha, beta and $\omega_h$ have the same values under the conditions with equal factor loadings, they are quite different under the conditions with unequal factor loadings. Among the three coefficients, beta has the lowest values while $\omega_h$ has the highest values; the discrepancy among them decreases as the degree of the average interitem correlation increases. All of the three coefficients are positively affected by
item number and the degree of the average interitem correlation. However, as discrepancy among the interitem correlations occurs, only beta decreases. The decrease in beta becomes moderate as the average interitem correlation increases from 0.3 to 0.6.

ECV is the only coefficient which does not change its value under all the unidimensional conditions, the values of which are expressed in proportions instead of percentages to serve the purpose of comparison. Under unidimensional conditions, the values of ECV are all ones, which indicates the perfect unidimensionality of a test. Although ECV works well for measuring the closeness to unidimensionality, it fails to reveal the change in the average interitem correlation and test internal structure.

**TABLE 1 Indices for measuring internal consistency for unidimensional data sets**

<table>
<thead>
<tr>
<th></th>
<th>Test length</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>10</td>
<td>20</td>
<td>40</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Equal Loadings</td>
<td>λᵢ = √0.3</td>
<td></td>
<td></td>
<td>λᵢ = √0.6</td>
<td></td>
</tr>
<tr>
<td>C’s ᵦᵢⱼ</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>SD</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>alpha</td>
<td>0.81</td>
<td>0.90</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>beta</td>
<td>0.81</td>
<td>0.90</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>ωₗ</td>
<td>0.81</td>
<td>0.90</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>ECV</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>Unequal Loadings</td>
<td>λᵢ = √0.62  and √0.10</td>
<td></td>
<td></td>
<td>λᵢ = √0.92  and √0.35</td>
<td></td>
</tr>
<tr>
<td>C’s ᵦᵢⱼ</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.60</td>
<td>0.60</td>
</tr>
<tr>
<td>SD</td>
<td>0.18</td>
<td>0.19</td>
<td>0.19</td>
<td>0.19</td>
<td>0.20</td>
</tr>
<tr>
<td>alpha</td>
<td>0.81</td>
<td>0.90</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>beta</td>
<td>0.68</td>
<td>0.74</td>
<td>0.78</td>
<td>0.89</td>
<td>0.92</td>
</tr>
<tr>
<td>ωₗ</td>
<td>0.83</td>
<td>0.90</td>
<td>0.95</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>ECV</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**Bifactor Data**

For bifactor data sets, the degree of general factor saturation is the main factor which affects the values of indices for measuring internal consistency. As the degree
of general factor saturation decreases while keeping other variables the same, all the indices except SD become smaller. Alpha was least influenced by the degree of general factor saturation while most influenced by test length. Beta and $\omega_h$ have the same values under all the conditions of bifactor data sets because the item loadings are equal in terms of the general factor, and each subscale has the same average within-subscale correlation. Beta and $\omega_h$ are affected dramatically by the degree of general factor loadings. As in Table 2, when the general factor loading decreases from $\sqrt{0.25}$ to $\sqrt{0.05}$, beta and $\omega_h$ drop by more than 70% of the values under the conditions with high general factor saturation. Although beta and $\omega_h$ are also affected by item number, the effect is not remarkable as alpha, especially under the conditions with low degree of general factor saturation. All the coefficients, except ECV, increase as the degree of average interitem correlation increases.

**TABLE 2 Indices for measuring internal consistency for bi-factor data sets**

<table>
<thead>
<tr>
<th>Subscale length</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>High GFS</td>
<td>$\lambda_i = \sqrt{0.25}$; $\alpha_{ik} = \sqrt{0.05}$</td>
<td>$\lambda_i = \sqrt{0.50}$; $\alpha_{ik} = \sqrt{0.10}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C's $\bar{r}_{ij}$</td>
<td>0.27</td>
<td>0.27</td>
<td>0.27</td>
<td>0.54</td>
<td>0.55</td>
<td>0.55</td>
</tr>
<tr>
<td>SD</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>alpha</td>
<td>0.79</td>
<td>0.88</td>
<td>0.94</td>
<td>0.92</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td>beta</td>
<td>0.72</td>
<td>0.81</td>
<td>0.85</td>
<td>0.85</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>$\omega_h$</td>
<td>0.72</td>
<td>0.81</td>
<td>0.85</td>
<td>0.85</td>
<td>0.88</td>
<td>0.89</td>
</tr>
<tr>
<td>ECV</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>Medium GFS</td>
<td>$\lambda_i = \sqrt{0.15}$; $\alpha_{ik} = \sqrt{0.15}$</td>
<td>$\lambda_i = \sqrt{0.30}$; $\alpha_{ik} = \sqrt{0.30}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C's $\bar{r}_{ij}$</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.43</td>
<td>0.44</td>
<td>0.45</td>
</tr>
<tr>
<td>SD</td>
<td>0.07</td>
<td>0.07</td>
<td>0.07</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>alpha</td>
<td>0.73</td>
<td>0.85</td>
<td>0.92</td>
<td>0.88</td>
<td>0.94</td>
<td>0.97</td>
</tr>
<tr>
<td>beta</td>
<td>0.51</td>
<td>0.58</td>
<td>0.62</td>
<td>0.61</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>$\omega_h$</td>
<td>0.51</td>
<td>0.58</td>
<td>0.62</td>
<td>0.61</td>
<td>0.64</td>
<td>0.65</td>
</tr>
<tr>
<td>ECV</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td>Low GFS</td>
<td>$\lambda_i = \sqrt{0.05}$; $\alpha_{ik} = \sqrt{0.25}$</td>
<td>$\lambda_i = \sqrt{0.10}$; $\alpha_{ik} = \sqrt{0.50}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C's $\bar{r}_{ij}$</td>
<td>0.16</td>
<td>0.17</td>
<td>0.17</td>
<td>0.32</td>
<td>0.34</td>
<td>0.34</td>
</tr>
<tr>
<td>SD</td>
<td>0.12</td>
<td>0.12</td>
<td>0.12</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>alpha</td>
<td>0.66</td>
<td>0.80</td>
<td>0.89</td>
<td>0.83</td>
<td>0.91</td>
<td>0.95</td>
</tr>
</tbody>
</table>
**Discussion and Recommendation**

To better evaluate which indices are appropriate for measuring internal consistency, we may review the factors which influence test internal consistency. 1) The degree of general factor saturation. Although all the coefficients are more or less influenced by that, only beta, $\omega_h$ and ECV can clearly indicate the change in the degree of general factor saturation. 2) The degree of the average interitem correlation. Except ECV and SD, all the other coefficients can reflect the change in the degree of the average interitem correlation. 3) Independence of test length. Alpha, beta and $\omega_h$ are all dependent of test length. They all increase with item number becoming larger while keeping other conditions the same. Alpha, as the estimate of reliability, is most influenced by test length among the three coefficients. Beta and $\omega_h$ are independent of item number under the condition that the general factor saturation is zero, which means there is no general factor but orthogonal group factors. Under that condition, beta and $\omega_h$ are always equal to zero. The indices independent of item number are Cronbach’s $\bar{r}_{ij}$, SD and ECV. Nevertheless, these indices, if used alone, fail to depict the whole picture of test internal consistency.

The results of the comparison of these indices show that measuring internal consistency with a statistic of a single value is not sufficient under the present stage, so a combination of measures for test internal consistency is recommended. To measure test internal consistency, we can first use ECV to assess whether a test is unidimensional or not, because it is the only index which can indicate the unidimensionality without the influence of other variables like test length or factor loading size. If ECV is one or close to one, then the test is unidimensional or very

<table>
<thead>
<tr>
<th></th>
<th>beta</th>
<th>$\omega_h$</th>
<th>ECV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.20</td>
<td>0.20</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.24</td>
<td>0.24</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.26</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.26</td>
<td>0.26</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.27</td>
<td>0.27</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>0.28</td>
<td>0.28</td>
<td>0.58</td>
</tr>
</tbody>
</table>
close to unidimensionality, the average degree of the correlations between items (Cronbach’s \( \bar{r}_{ij} \)) in the whole test and the standard deviation of these interitem correlations should be both reported for indicating test internal consistency. If ECV is rather less than 1 (e.g., 0.75 or below in this study), then beta and \( \omega_h \) are recommended because they are more affected by the degree of general factor saturation than by the degree of average interitem correlations while Cronbach’s \( \bar{r}_{ij} \) is the other way around; besides, beta and \( \omega_h \) are less affected by test length when ECV is 0.75 or below. Under the more stringent definition of internal consistency, only beta is recommended when ECV is rather less than 1 because beta decreases as discrepancy among the interitem correlations increases while \( \omega_h \) does not indicate the disagreement among the interitem correlations.

To distinguish internal consistency from other terms in the fields of measurement and testing, explicit and less confusing names for different properties of a test are also needed. For example, the term “internal consistency reliability” is confusing and it can be replaced with reliability estimates under the condition of internal consistency. In short, more efforts are required for the clarification of different psychometrical terms so that appropriate measures can be developed and applied in the future.

References


Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach’s $\alpha$, Revelle’s $\beta$, and McDonald’s $\omega_H$: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*, 123-133.