PSYCHOMETRIKA—VOL. 74, NO. 1, 145–154
MARCH 2009
DOI: 10.1007/S11336-008-9102-Z

COEFFICIENTS ALPHA, BETA, OMEGA, AND THE GLB: COMMENTS ON SIJTSMA

WILLIAM REVELLE

DEPARTMENT OF PSYCHOLOGY, NORTHWESTERN UNIVERSITY

RICHARD E. ZINBARG

DEPARTMENT OF PSYCHOLOGY, THE FAMILY INSTITUTE AT NORTHWESTERN UNIVERSITY, NORTHWESTERN UNIVERSITY

There are three fundamental problems in Sijtsma (Psychometrika, 2008): (1) contrary to the name, the glb is not the greatest lower bound of reliability but rather is systematically less than ω_t (McDonald, Test theory: A unified treatment, Erlbaum, Hillsdale, 1999), (2) we agree with Sijtsma that when considering how well a test measures one concept, α is not appropriate, but recommend ω_t rather than the glb, and (3) the end user needs procedures that are readily available in open source software.

Key words: reliability, internal consistency, homogeneity, test theory, coefficient alpha, coefficient omega, coefficient beta.

The problem of how to assess reliability has been with us ever since Spearman (1904) introduced the concept of correction for attenuation and that of split half reliability (Spearman, 1910). To Spearman (1904), reliability was used as a way of finding the "real correlation between the true objective values" (r_{pq}) by correcting observed correlations ($r_{p'q'}$) for the attenuation of "accidental" deviations of observed scores from their "true objective values." To Spearman (1904, p. 90), this required finding "the average correlation between one and another of these independently obtained series of values" (what has come to be called parallel tests) to estimate the reliability of each set of measures ($r_{p'p'}, r_{q'q'}$), and then to find

$$r_{pq} = \frac{r_{p'q'}}{\sqrt{r_{p'p'} r_{q'q'}}}. \tag{1}$$

Rephrasing Spearman (1904, 1910) in more current terminology (Lord & Novick, 1968; McDonald, 1999), reliability is the correlation between two parallel tests where tests are said to be parallel if for every subject, the true scores on each test are the expected scores across an infinite number of tests, and thus the same, and the error variances across subjects for each test are the same. Unfortunately, "all measurement is befuddled by error" (McNemar, 1946, p. 294). Error may be defined as observed score − true score, and hence to be uncorrelated with true score and uncorrelated across tests. Thus, reliability is the fraction of test variance that is true score variance. However, such a definition requires finding a parallel test. For just knowing the correlation between two tests, without knowing the true scores or their variance (and if we did, we would not bother with reliability), we are faced with three knowns (two variances and one covariance), but ten unknowns (four variances and six covariances).

In this case of two tests, by defining them to be parallel with uncorrelated errors, the number of unknowns drops to three and reliability of each test may be found. With three tests, the number of assumptions may be reduced, and if the tests are tau (τ) equivalent (each test has the same true score covariance), reliability for each of the three tests may be found. With four tests, to find

Requests for reprints should be sent to William Revelle, Department of Psychology, Northwestern University, Evanston, IL, USA. E-mail: revelle@northwestern.edu

the reliability of each test, we need only assume that the tests all measure the same construct (to be "congeneric"), although possibly with different true score and error score variances (Lord & Novick, 1968).

Unfortunately, with rare exceptions, we normally are faced with just one test, not two, three, or four. How then to estimate the reliability of that one test? The original solution was to estimate reliability based upon the correlation between two halves (r_1) correcting for the fact they were half tests rather than full tests using a special case ($n = 2$) of the more general Spearman–Brown correction (Brown, 1910; Spearman, 1910)

$$r_{xx} = \frac{nr_1}{1 + (n-1)r_1}. \tag{2}$$

Subsequent efforts were based on the domain sampling model in which tests are seen as being made up of items randomly sampled from a domain of items (Lord, 1955, made the distinction between "Type 1" sampling of people, "Type 2" sampling of items, and "Type 12" sampling of persons and items). The desire for an easy to use "magic bullet" based upon the domain sampling model has led to a number of solutions (e.g., the six considered by Guttman, 1945), of which one, coefficient *alpha* (Cronbach, 1951) is easy to compute and easy to understand. The appeal of α was perhaps that it was the average of all such random splits (Cronbach, 1951).

Even though the pages of *Psychometrika* have been filled over the years with critiques and cautions about coefficient α and have seen elegant solutions for more appropriate estimates, few of these suggested coefficients are used. This is partly because they are not easily available in programs for the end user nor described in a language that is accessible to many psychologists. In a statement reminiscent of Spearman's observation that "Psychologists, with scarely an exception, never seem to have become acquainted with the brilliant work being carried on since 1886 by the Galton–Pearson school" (Spearman, 1904, p. 96), Sijtsma (2008) points out that psychometrics and psychology have drifted apart as psychometrics has become more statistical and psychologists have remained psychologists. Without clear discussions of the alternatives and easily available programs to find the alternative estimates of reliability, most psychologists will continue to use α. With the advent of open source programming environments for statistics such as R (R Development Core Team, 2008), that are easy to access and straightforward to use, it is possible that the other estimates of reliability will become more commonly used.

What coefficients should we use? Sijtsma (2008) reviews a hierarchy of lower bound estimates of reliability and in agreement with Jackson and Agunwamba (1977) and Woodhouse and Jackson (1977) suggests that the glb or "greatest lower bound" (Bentler & Woodward, 1980) is, in fact, the best estimate. We believe that this is an inappropriate suggestion for at least three reasons:

1. Contrary to what the name implies, the glb is not the greatest lower bound estimate of reliability, but is somewhat less than another, easily calculated and understood estimate of reliability (ω_{total}, ω_t) of McDonald (1999). (We use the subscript on ω_t to distinguish between the coefficient ω introduced by McDonald (1978), equation (9), and McDonald (1999), equation (6.20) that he also called ω and which we (Zinbarg, Revelle, & Yovel, 2005) previously relabeled $\omega_{\text{hierarchical}}$, ω_h).
2. Rather than just focusing on the greatest lower bounds as estimates of a reliability of a test, we should also be concerned with the percentage of the test that measures one construct. As has been discussed previously (Revelle, 1979; McDonald, 1999; Zinbarg et al., 2005), this may be estimated by finding ω_h, the general factor saturation of the test (McDonald, 1999; Zinbarg et al., 2005), or the worst split half reliability of a test (coefficient beta, β, of Revelle, 1979).
3. Although it is easy to estimate all of the Guttman (1945) lower bounds, as well as β, ω_h, and ω_t, the techniques for estimating the glb are not readily available for the end user.

1. The Ordering of Reliability Estimates

Defined as the correlation between a test and a test just like it, reliability would seem to require a second test. The traditional solution when faced with just one test is to consider the internal structure of that test. Letting reliability be the ratio of true score variance to test score variance, or alternatively, 1 − the ratio of error variance to true score variance, the problem becomes one of estimating the amount of error variance in the test. That is, two tests, $\mathbf{X}$, and a test just like it, $\mathbf{X'}$, with covariance, $\mathbf{C_{xx'}}$ can be represented as

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_x} & \vdots & \mathbf{C_{xx'}} \\ \cdots & \cdots & \cdots \\ \mathbf{C_{xx'}} & \vdots & \mathbf{V_{x'}} \end{pmatrix} \tag{3}$$

and letting $V_{\mathbf{x}} = \mathbf{1V_x1'}$ and $C_{\mathbf{XX'}} = \mathbf{1C}_{XX'}\mathbf{1'}$ the correlation between the two tests will be

$$\rho = \frac{C_{xx'}}{\sqrt{V_x V_{x'}}}. \tag{4}$$

Although arguing that reliability was only meaningful in the case of test-retest, Guttman (1945) may be credited with introducing a series of lower bounds for reliability, each based upon the item characteristics of a single test. These six have formed the base for most of the subsequent estimates.

All of these estimates assume that the covariances between items represent true covariance, but that the variances of the items reflect an unknown sum of true and unique variance. That is, the variance of a test is simply the sum of the true covariances and the error variances:

$$V_{\mathbf{x}} = \mathbf{1V_x1'} = \mathbf{1C_t1'} + \mathbf{1V_e1'} = V_t + V_e \tag{5}$$

and the structure of the two tests seen in (3) becomes

$$\Sigma_{XX'} = \begin{pmatrix} \mathbf{V_X = V_t + V_e} & \vdots & \mathbf{C_{xx'} = V_t} \\ \cdots & \cdots & \cdots \\ \mathbf{V_t = C_{xx'}} & \vdots & \mathbf{V_{t'} + V_{e'} = V}_{X'} \end{pmatrix} \tag{6}$$

and because $\mathbf{V}_t = \mathbf{V}_{t'}$ and $\mathbf{V}_e = \mathbf{V}_{e'}$ reliability is

$$\rho = \frac{C_{XX'}}{V_X} = \frac{V_t}{V_X} = 1 - \frac{V_e}{V_t}. \tag{7}$$

The problem remains how to estimate V_t and V_e from one test. Guttman (1945), in an attempt to formalize the estimation of reliability, proposed six lower bounds for ρ. Each one successively modifies the way that the error variance of the items are estimated. The first lowest bound, λ_1 considers that all of an item variance is error and that only the interitem covariances reflect true variability. Thus, λ_1 subtracts the sum of the diagonal of the observed item covariance matrix from the total test variance:

$$\lambda_1 = 1 - \frac{\mathrm{tr}(\mathbf{V_x})}{V_x} = \frac{V_x - \mathrm{tr}(\mathbf{V}_x)}{V_x}. \tag{8}$$

The second bound, λ_2 replaces the diagonal with a function of the square root of the sums of squares of the off diagonal elements. Let $C_2 = \mathbf{1}(\mathbf{V} - \text{diag}(\mathbf{V}))^2\mathbf{1}'$, then

$$\lambda_2 = \lambda_1 + \frac{\sqrt{\frac{n}{n-1}C_2}}{V_x} = \frac{V_x - \text{tr}(\mathbf{V}_x) + \sqrt{\frac{n}{n-1}C_2}}{V_x}. \tag{9}$$

Effectively, this is replacing the diagonal with n^* the square root of the average squared off diagonal element.

Guttman's third lower bound, λ_3, also modifies λ_1 and estimates the true variance of each item as the average covariance between items and is, of course, the same as Cronbach's α

$$\lambda_3 = \lambda_1 + \frac{\frac{V_X - \text{tr}(\mathbf{V}_X)}{n(n-1)}}{V_X} = \frac{n\lambda_1}{n-1} = \frac{n}{n-1}\left(1 - \frac{\text{tr}(\mathbf{V})_x}{V_x}\right) = \frac{n}{n-1}\frac{V_x - \text{tr}(\mathbf{V}_x)}{V_x} = \alpha. \tag{10}$$

This is just replacing the diagonal elements with the average off diagonal elements. $\lambda_2 \geq \lambda_3$ with $\lambda_2 > \lambda_3$ if the covariances are not identical.

As pointed out by Ten Berge and Zegers (1978), λ_3 and λ_2 are both corrections to λ_1 and this correction may be generalized as an infinite set of successive improvements

$$\mu_r = \frac{1}{V_x}\left(p_o + \left(p_1 + \left(p_2 + \cdots \left(p_{r-1} + (p_r)^{1/2}\right)^{1/2} \cdots\right)^{1/2}\right)^{1/2}\right), \quad r = 0, 1, 2, \ldots \tag{11}$$

where

$$p_h = \sum_{i \neq j} \sigma_{ij}^{2h}, \quad h = 0, 1, 2, \ldots, r-1$$

and

$$p_h = \frac{n}{n-1}\sigma_{ij}^{2h}, \quad h = r$$

(Ten Berge & Zegers, 1978). Clearly, $\mu_0 = \lambda_3 = \alpha$ and $\mu_1 = \lambda_2$. $\mu_r \geq \mu_{r-1} \geq \cdots \geq \mu_1 \geq \mu_0$, although the series does not improve much after the first two steps.

Guttman's fourth lower bound, λ_4, was originally proposed as any spit half reliability (Guttman, 1945), but has been interpreted as the greatest split half reliability (Jackson & Agunwamba, 1977). If $\mathbf{X}$ is split into two parts, $\mathbf{X}_a$ and $\mathbf{X}_b$, with correlation r_{ab} then

$$\lambda_4 = 2\left(1 - \frac{V_{X_a} + V_{X_b}}{V_X}\right) = \frac{4r_{ab}}{V_x} = \frac{4r_{ab}}{V_{X_a} + V_{X_b} + 2r_{ab}V_{X_a}V_{X_b}} \tag{12}$$

which is just the normal split half reliability, but in this case, of the most similar splits.

λ_5, Guttman's fifth lower bound, replaces the diagonal values with twice the square root of the maximum (across items) of the sums of squared interitem covariances

$$\lambda_5 = \lambda_1 + \frac{2\sqrt{\bar{C}_2}}{V_X}. \tag{13}$$

Although superior to λ_1, λ_5 underestimates the correction to the diagonal. A better estimate would be analogous to the correction used in λ_3:

$$\lambda_{5+} = \lambda_1 + \frac{n}{n-1}\frac{2\sqrt{\bar{C}_2}}{V_X}. \tag{14}$$

Guttman's final bound considers the amount of variance in each item that can be accounted for by the linear regression of all of the other items (the squared multiple correlation or smc), or more precisely, the variance of the errors, e_j^2, and is

$$\lambda_6 = 1 - \frac{\sum e_j^2}{V_x} = 1 - \frac{\sum(1 - r_{\text{smc}}^2)}{V_x}. \tag{15}$$

Not included in Guttman's list of lower bounds is McDonald's ω_t, which is similar to λ_6, but uses the estimates of uniqueness (u^2) from factor analysis to find e_j^2. This is based on a decomposition of the variance of a test score, V_x, into four parts: that due to a general factor, **g**, that due to a set of group factors, **f** (factors common to some but not all of the items), specific factors, **s** unique to each item, and **e**, random error. (Because specific variance can not be distinguished from random error unless the test is given at least twice, McDonald (1999) combines these both into error). Letting

$$\mathbf{x} = \mathbf{cg} + \mathbf{Af} + \mathbf{Ds} + \mathbf{e} \tag{16}$$

then the communality of item$_j$, based upon general as well as group factors,

$$h_j^2 = c_j^2 + \sum f_{ij}^2 \tag{17}$$

and the unique variance for the item

$$u_j^2 = \sigma_j^2\left(1 - h_j^2\right) \tag{18}$$

may be used to estimate the test reliability. That is, if h_j^2 is the communality of item$_j$, based upon general as well as group factors, then for standardized items, $e_j^2 = 1 - h_j^2$ and

$$\omega_t = \frac{\mathbf{1cc'1} + \mathbf{1AA'1}'}{V_x} = 1 - \frac{\sum(1 - h_j^2)}{V_x} = 1 - \frac{\sum u^2}{V_x}. \tag{19}$$

Because $h_j^2 \geq r_{\text{smc}}^2$, $\omega_t \geq \lambda_6$.

It is important to distinguish here between the two ω coefficients of McDonald (1978) and (McDonald, 1999, equation (6.20a)), ω_t and ω_h. While the former is based upon the sum of squared loadings on all the factors, the latter is based upon the sum of the squared loadings on the general factor

$$\omega_h = \frac{\mathbf{1cc'1}}{V_x}. \tag{20}$$

As we will discuss later, ω_h is a very important indicator of how well a test measures one construct.

Yet another estimate that has been proposed for the reliability of a principal component (Ten Berge & Hofstee, 1999) unfortunately also uses λ_1 as a symbol, but this time as the magnitude of the first eigenvalue is

$$\alpha_{\text{pc}} = 1 - \frac{n}{(n-1)\lambda_1}. \tag{21}$$

The discussion of various lower bounds seemed finished when Jackson and Agunwamba (1977) and Bentler and Woodward (1980) introduced their "greatest lower bound," or glb. Woodhouse and Jackson (1977) organized Guttman's six bounds into a series of partial orders, and provided an algorithm for estimating the glb of Jackson and Agunwamba (1977). An alternative

algorithm was proposed by Bentler and Woodward (1980) and discussed by Sijtsma (2008). Unfortunately, none of these authors considered ω_t, which we will show tends to exceed the glbs reported in the various discussions of the utility of the glb.

2. A Comparison of Estimates of Reliability: When Is the Greatest Lower Bound Not the Greatest?

To understand how Guttman's bounds relate to each other and to the glb and ω_t, it is useful to consider some now classic example data sets. Using open source functions available in the **psych** package (Revelle, 2008) for R (R Development Core Team, 2008), we compared the six lower bounds of Guttman (1945), two ways of estimating α, one of which uses the traditional approach, ($\lambda_3 = \alpha$), the second of which is the α of the first principal component, four of the bounds of Ten Berge and Zegers (1978), the glb, and ω_t for nine data sets (Table 1). The first six are taken from Sijtsma (2008) who reports three real and three artificial examples. We also examine two of those of Bentler and Woodward (1980) who report examples of their algorithm for computing the glb. The first set was taken from Lord and Woodward (1968), the second from Warner (1960). The final comparison is from Ten Berge and Socan (2004) who gives an example taken from De Leeuw (1983). Two other estimates reported in Table 1 that will be discussed later are coefficients β (Revelle, 1979) and ω_h (McDonald, 1999; Zinbarg et al., 2005). Although Confirmatory Factor Analysis (CFA) or Structural Equation Modeling (SEM) techniques could have been used to estimate ω_t (Raykov & Shrout, 2002) and ω_h (Zinbarg, Yovel, Revelle & McDonald, 2006, 2007), we made use of Exploratory Factor Analysis (EFA).

Two findings are very clear from Table 1: α is much lower than the superior estimates of reliability and the highest estimate of reliability is never the glb. In most, but not all of the examples, ω_t (McDonald's estimate of the proportion of total common variance in the test) provides the greatest reliability estimate. The two exceptions are when the maximum split half reliability is the greatest reliability. The differences between the three highest estimates (ω_t, λ_4 and the glb)

TABLE 1.
Comparison of 13 estimates of reliability. The data sets and the glb estimates are taken from the six examples in Sijtsma (2008) (S1–S2c), two examples in Bentler and Woodward (1980), (B&W 1 & 2) and the De Leeuw (1983) dataset analyzed by Ten Berge and Socan (2004). The greatest reliability estimates are underlined. $\lambda_1 \ldots \lambda_6$ are the Guttman (1945) bounds, ω_h and ω_t are from McDonald (1999), $\mu_0 \ldots \mu_3$ are from Ten Berge and Zegers (1978), β is from Revelle (1979).

Estimate	S-1	S-1a	S-1b	S-2a	S-2b	S-2c	B&W 1	B&W 2	TB& S
N items	8	4	4	6	6	6	4	6	6
β (min)	.656	.651	.610	.000	.000	.437	.756	.854	.739
ω_h	.593	.643	.676	.049	.000	.532	.706	.921	.767
λ_1	.687	.561	.507	.444	.444	.444	.671	.785	.700
$\lambda_3(\alpha, \mu_0)$	.785	.749	.676	.533	.533	.533	.894	.942	.840
α_{pc}	.787	.749	.676	.553	.533	.553	.896	.943	.841
$\lambda_2(\mu_1)$	.789	.753	.678	.643	.585	.533	.898	.943	.842
μ_2	.790	.755	.657	.663	.592	.533	.899	.943	.843
μ_3	.791	.755	.658	.666	.592	.533	.900	.943	.843
λ_5	.766	.738	.660	.593	.549	.511	.881	.911	.819
λ_6 (smc)	.785	.713	.593	.800	.571	.488	.880	.960	.830
λ_4 (max)	<u>.853</u>	.820	.696	.889	.647	.533	.913	<u>.979</u>	.884
glb	.852	.820	.696	.889	.667	.533	.920	.976	.885
ω_t	.844	<u>.893</u>	<u>.859</u>	<u>.889</u>	<u>.669</u>	<u>.561</u>	<u>.951</u>	.972	<u>.900</u>

tend to be not great (indeed several are only observable at the third decimal point) and all three differ substantially from α.

In that reliability is used to correct for attenuation (equation (1)), underestimating the reliability will lead to an overestimate of the unattenuated correlation and overestimating the reliability will lead to an underestimate of the unattenuated correlation. Choosing the proper reliability coefficient is therefore very important and should be guided by careful thought and strong theory. In the case in which our test is multidimensional and several of the dimensions contribute to the prediction of the criterion of interest, α will lead to an overcorrection, but unfortunately, so will using the glb. ω_t will lead to a more accurate correction. In the case in which the test is multidimensional, but only the test's general factor contributes to the prediction of the criterion of interest, α will lead to an undercorrection, and the glb will unfortunately lead to an even greater undercorrection of the estimate of the association between the test's general factor and the criterion. ω_h would lead to a more accurate correction in this case.

3. What Is the Meaning of Internal Consistency?

We agree with Sijtsma (2008) and indeed have long argued that α is a poor index of unidimensionality (Revelle, 1979; Zinbarg et al., 2005); so, in fact, did Cronbach (1951, 1988); Cronbach and Shavelson (2004). The examples of varying the factor structure from one to three factors while maintaining an equal α (Sijtsma, 2008, Table 5) are helpful, for they show the insensitivity to internal structure of some of the Guttman (1945) indices. However, rather than using the amount of explained common variance (ECV) suggested by Sijtsma (2008), we believe that a more appropriate measure to consider is an index of how much the test measures *one common factor*. We reiterate our recommendation (Zinbarg et al., 2005, 2006) for use of higher factor analysis with a Schmid–Leiman transformation (Schmid & Leiman, 1957) (if doing EFA) and the subsequent estimation of the *general factor* saturation (coefficient ω_h of McDonald, 1999, equation (6.21)). This may also be done, of course, using SEM or CFA procedures for hierarchical factor analysis. Alternatively, at least one of us also likes to use hierarchical cluster analysis of the items to find the worst split half reliability (coefficient β of Revelle, 1979). Both of these coefficients are estimates of the amount of variance attributable to one common factor for all of the items.[1] It is particularly telling that the β and ω_h estimates are 0.0[2] for the two examples of Sijtsma (2008) of data with multiple factors (Table 1). Even though $\alpha = .533$ and the glb and ω_t were very large, and thus show reliability in the sense of relatively little error variability; they do not show homogeneity or internal consistency. The ECV estimate preferred by Sijtsma (2008) does show that a test is not unidimensional, but we find that the 50% and 33% ECV values not as compelling as the 0.0 values for β or ω_h.

The issue of how much information is available in a single testing session is very important. When using tests meant to assess individual differences in stable traits such as verbal ability or neuroticism, the idea of reliability defined as stability across time makes sense. Indeed, by using multiple measures we are able to distinguish between unique, but reliable versus unique and unreliable item variance. But when evaluating how well we are assessing levels of particular states, such as energetic arousal or positive affect, at a particular time (Rafaeli & Revelle, 2006), we must use some estimate of internal consistency.

As for the meaning of internal consistency, we agree with Sijtsma (2008) that it has been used in different ways by different authors. Some have used it to be synonymous with homogeneity and unidimensionality. Others have reserved homogeneity to refer to unidimensionality

[1] But see Zinbarg et al. (2005) for a discussion of why ω_h might be preferred.

[2] As the careful reader will note, using EFA the estimated ω_h in set S2a was .04 rather than 0.0. Using CFA this becomes 0.0.

and internal consistency to refer to interrelatedness of items. The problem with interrelatedness is that it does not differentiate between the case in which each item is related to only a small proportion of the other items in the test from the case in which each item is related to every or nearly every other item in the test.

In our view, there are four important psychometric properties that a test might possess.

1. Unidimensionality, as in IRT, whether a single latent variable is being measured (thus, we could imagine a case—in which difficulty factors are present—in which a unidimensional scale has a factor structure that has multiple factors when represented in terms of the linear factor model that does not do a very good job when items have nonlinear associations with the latent variable under at least some conditions). Whereas unidimensionality represents the ideal of measurement (McNemar, 1946), there are some domains that consist of related (possibly even highly related) yet discriminable facets that differ even in their true scores. Such domains are themselves not unidimensional, and so it would be unrealistic to expect measures of them to be unidimensional.
2. Presence of a general factor. If all of the facets in a domain are related, at least to some extent, then there is a single latent variable that is common to all of the items in that domain. For those cases in which the ideal of unidimensionality is not realistic (see above), the presence of a general factor is the ideal for which to strive. The presence of a general factor can be tested via the appropriate testing of nested confirmatory factor models (i.e., comparing a model with k orthogonal group factors at the highest level of its factor structure to either a (1) model with k orthogonal group factors at the highest level of its factor structure that also contains a general factor orthogonal to all other factors or (2) a model with an additional level to its structure that is loaded on by the k group factors that form the highest level of the comparison model).
3. The proportion of test variance due to a general factor. It is possible that a test is unidimensional or contains a general factor, but that factor common to all of the test items is so weakly saturated in the items that the test does not provide a precise measure of the single latent variable or general factor. Thus, the proportion of test variance due to a general factor provides important information because it indexes the precision with which the test's total scores estimate a latent variable common to all test items (alternatively, this proportion indexes the degree to which the total scores generalize to latent variable common to all test items). This proportion is ω_h.
4. The proportion of test variance due to all common factors. There are some contexts, such as applied prediction, in which we are concerned with the upper bound of the extent to which a test's total score can correlate with some other measure and we are not concerned with theoretical understanding regarding which constructs are responsible for that correlation. The proportion of test variance due to all common factors provides this upper bound (alternatively, this proportion indexes generalizability to the domain from which the test items are a representative sample and which may represent more than one latent variable). This proportion is ω_t (and which $= \alpha$ when the test is unidimensional).

If people insist on continuing to use the terms homogeneity and internal consistency, perhaps they would use the labels unidimensionality for property 1, homogeneity for property 2 (presence of a general factor, the items are homogeneous to the extent that they all share at least one attribute or latent variable common), general factor saturation for property 3, and internal consistency for property 4.

4. Estimation of Reliability

It has been known for a long time that α is a lower bound to the reliability, in many cases even a gross underestimate, and a poor estimate of internal consistency and in some cases a gross overestimate, but it continues to be used. Why is this? Perhaps inertia on the part of editors and reviewers who insist on at least some estimate of reliability and do not know what to recommend. Perhaps inertia on the part of commercial program to implement features that are not widely requested. And perhaps it is the fault of psychometricians who develop better and more powerful algorithms, but do not make them readily available. A case in point is the Minimum Rank Factor Analysis program used for some of the examples in Sijtsma (2008). It is said to be available from the web, but it turns out to be a Pascal program that runs just on MS-DOS. This is not overly helpful for users of non-MS-DOS platforms. With the wide acceptance of open source programming systems such as R (R Development Core Team, 2008) that run on all platforms, perhaps it is time to implement the better estimates in open source programs. We have done so with the implementation of ω_h, ω_t, β and $\lambda_1 \ldots \lambda_6$ in the **psych** package available from CRAN (the Comprehensive R Archive Network: http://www.R-project.org). (For examples of syntax for estimating ω_t using proprietary SEM software, see Raykov and Shrout, 2002.) We encourage others to do the same.

5. Conclusions

We concur with Sijtsma (2008) that editors and authors should be encouraged to report better estimates of reliability in addition to α. Where we disagree is what estimates to report. The recommendation for using the glb as the best estimate of reliability is more of a marketing ploy based upon the name of "greatest lower bound" rather than reality. As is clear from Table 1, McDonald's ω_t exceeds the glb in all but two of the examples given by either Sijtsma (2008), Bentler and Woodward (1980) or Ten Berge and Socan (2004). In those two cases, the maximum split half reliability slightly exceeds the glb. We have previously discussed the many situations where it is very important to estimate ω_h (Zinbarg et al., 2005, 2006, 2007). ω_t and ω_h are easy to calculate from any factor analysis output in either commercial programs (e.g., SPSS or SAS) or packages (e.g., **psych**, Revelle, 2008) contributed to the open source program R (R Development Core Team, 2008). It is likely that psychometric contributions would have greater impact if they were readily available in such open source programs.

References

Bentler, P., & Woodward, J. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*(2), 249–267.

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, *3*(3), 296–322.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

Cronbach, L.J. (1988). Internal consistency of tests: Analyses old and new. *Psychometrika*, *53*(1), 63–70.

Cronbach, L.J., & Shavelson, R.J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*(3), 391–418.

De Leeuw, J. (1983). Models and methods for the analysis of correlation coefficients. *Journal of Econometrics*, *22*(1–2), 113–137.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*(4), 255–282.

Jackson, P., & Agunwamba, C. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: I: Algebraic lower bounds. *Psychometrika*, *42*(4), 567–578.

Lord, F.M. (1955). Sampling fluctuations resulting from the sampling of test items. *Psychometrika*, *20*(1), 1–22.

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.

McDonald, R.P. (1978). Generalizability in factorable domains: "domain validity and generalizability": 1. *Educational and Psychological Measurement*, *38*(1), 75–79.

McDonald, R.P. (1999). *Test theory: A unified treatment*. Hillsdale: Erlbaum.

McNemar, Q. (1946). Opinion-attitude methodology. *Psychological Bulletin*, *43*(4), 289–374.
R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria (ISBN 3-900051-07-0).
Rafaeli, E., & Revelle, W. (2006). A premature consensus: Are happiness and sadness truly opposite affects? *Motivation and Emotion*, *30*(1), 1–12.
Raykov, T., & Shrout, P.E. (2002). Reliability of scales with general structure: Point and interval estimation using a structural equation modeling approach. *Structural Equation Modeling*, *9*(2), 195–212.
Revelle, W. (1979). Hierarchical cluster-analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74.
Revelle, W. (2008). *psych: Procedures for personality and psychological research* (R package version 1.0-51).
Schmid, J.J., & Leiman, J.M. (1957). The development of hierarchical factor solutions. *Psychometrika*, *22*(1), 83–90.
Sijtsma, K. (2008). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*.
Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*(1), 72–101.
Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, *3*(3), 271–295.
Ten Berge, J.M.F., & Hofstee, W.K.B. (1999). Coefficients alpha and reliabilities of unrotated and rotated components. *Psychometrika*, *64*(1), 83–90.
Ten Berge, J.M.F., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*(4), 613–625.
Ten Berge, J.M.F., & Zegers, F.E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*(4), 575–579.
Warner, W.L. (1960). *Social class in America: a manual of procedure for the measurement of social status*. New York: Harper.
Woodhouse, B., & Jackson, P. (1977). Lower bounds for the reliability of the total score on a test composed of non-homogeneous items: II: A search procedure to locate the greatest lower bound. *Psychometrika*, *42*(4), 579–591.
Zinbarg, R.E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ω_H: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133.
Zinbarg, R.E., Yovel, I., Revelle, W., & McDonald, R.P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement*, *30*(2), 121–144.
Zinbarg, R.E., Revelle, W., & Yovel, I. (2007). Estimating ω_h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement*, *31*(2), 135–157.

Published Online Date: 11 DEC 2008