PSYCHOMETRIKA—VOL. 74, NO. 1, 107–120
MARCH 2009
DOI: 10.1007/S11336-008-9101-0

ON THE USE, THE MISUSE, AND THE VERY LIMITED USEFULNESS OF CRONBACH'S ALPHA

KLAAS SIJTSMA

TILBURG UNIVERSITY

This discussion paper argues that both the use of Cronbach's alpha as a reliability estimate and as a measure of internal consistency suffer from major problems. First, alpha always has a value, which cannot be equal to the test score's reliability given the interitem covariance matrix and the usual assumptions about measurement error. Second, in practice, alpha is used more often as a measure of the test's internal consistency than as an estimate of reliability. However, it can be shown easily that alpha is unrelated to the internal structure of the test. It is further discussed that statistics based on a single test administration do not convey much information about the accuracy of individuals' test performance. The paper ends with a list of conclusions about the usefulness of alpha.

Key words: Cronbach's alpha, internal consistency, reliability, unidimensionality.

1. Introduction

Probably no other statistic has been reported more often as a quality indicator of test scores than Cronbach's (1951) alpha coefficient, and presumably no other statistic has been subject to so much misunderstanding and confusion. Two problems concerning alpha continue to be pervasive in test construction and test use. The first problem is twofold. Alpha is a lower bound to the reliability, in many cases, even a gross underestimate, and alpha cannot have a value that could be the reliability based on the usual assumptions about measurement error. Better alternatives to alpha exist but are hardly known, let alone used to assess reliability. Thus, by continuing to use alpha as *the* estimate of reliability test constructors and test users do themselves injustice until they recognize the availability of better alternatives. The second problem is that alpha is persistently and incorrectly taken to be a measure of the internal structure of the test, and hence as evidence that the items in the test "measure the same thing." However, alpha does not provide the researcher with this sort of information. The result of this misinterpretation of alpha is that due to a high alpha value, trait validity (Campbell, 1960) often is taken for granted when, in fact, it has not been investigated at all.

Because alpha continues to be so important, in particular to practical researchers, and because alpha continues to be the subject of so much misinterpretation, there appears to be a strong need to settle some issues and provide suggestions for the practical estimation of test score reliability and the assessment of what the test measures. The goal of this paper is to illuminate the flaws and fallacies that surround both the "common" knowledge base and the practical use of Cronbach's alpha, and to provide alternatives. This paper is also meant to invite debate on topics that psychometricians often seem to overlook and test constructors and test practitioners tend to take for granted. The paper only uses knowledge that has been around for a while, but somehow has failed to come through well enough.

The paper is organized as follows. First, some historical facts about Cronbach's alpha are discussed. Second, the definitions of test score reliability, the greatest lower bound (glb; e.g.,

Requests for reprints should be sent to Klaas Sijtsma, Department of Methodology and Statistics, Faculty of Social Sciences, Tilburg University, PO Box 90153, 5000LE Tilburg, The Netherlands. E-mail: k.sijtsma@uvt.nl

© 2008 The Author(s). This article is published with open access at Springerlink.com

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

Woodhouse & Jackson, 1977) to the reliability, and alpha are discussed, and the relationships between alpha, the glb, and the reliability are outlined. Third, the application to real data of alpha, the *greater* lower bound Guttman's (1945) λ_2, and the glb is discussed, and it is shown that also in real data both alpha and λ_2 can be considerably smaller than the glb. Fourth, it is explained how alpha came to be misunderstood as a measure of internal consistency and it is shown that, in general, alpha does not convey information on the internal structure of the test. Fifth, it is argued that reliability estimates based on a single test administration, like alpha, may not convey much information about the accuracy of individual test performance. This contribution ends with five conclusions about the usefulness of alpha and alternative reliability estimation procedures.

2. Historical Facts

As happens so often, great inventions do not carry the name of their inventor, but instead of the researcher who was most successful in outlining its favorable properties in such a way that all of a sudden everything seemed to fall into place. It is no different with Cronbach's alpha. To avoid misunderstandings, Cronbach (1951) himself did not claim alpha to be his invention, but at great length credited results with respect to alpha to other authors. These authors include Kuder and Richardson (1937), who published a version of alpha for dichotomous items that went under the name of KR20. Another author is Hoyt (1941), who proposed a method for estimating reliability based on an analysis-of-variance decomposition of the data, which for dichotomous items gives the same results as KR20. Guttman (1945) derived alpha, denoted by the indexed Greek lower case λ_3, as the third in a series of six coefficients each of which was shown to be a lower bound to the reliability (also, see Jackson & Agunwamba, 1977). The derivation of alpha and the other coefficients used continuous random variables for item scores, and thus include dichotomous and ordered polytomous scoring as special cases. For dichotomously scored items, KR20, and to some extent a computationally convenient approximation denoted KR21—notice there were no computers in those days—actually gained quite some fame, but they were gradually pushed aside as alpha conquered territory.

Ever since its publication in 1951 in *Psychometrika*, Cronbach's famous paper has been a landmark to psychometricians, test constructors and test practitioners. Today, the paper still is one of the most downloaded papers from *Psychometrika*'s website (accessible via http://www.springer.com). Web of Science reports over 6,500 citations, which is a crushing number compared to the already very respectable 400+ citations for Kuder and Richardson (1937) and 200+ citations for Guttman (1945). For what it is worth (perhaps biological articles become outdated quicker than psychometric articles), it even outranks Watson and Crick's famous 1953 *Nature* article in which they describe the discovery of the double helix structure of DNA. Almost no psychological test or inventory is published without alpha being reported (usually without reference to Cronbach's paper), often for each interesting subgroup separately. Also, alpha continues to receive interest in psychometric research. For example, Van Zyl, Neudecker, and Nel (2000), Kistner and Muller (2004), and Hayashi and Kamata (2005) in different ways addressed the distribution of alpha; Ten Berge and Sočan (2004) discussed the relationship between alpha and other lower bounds and test unidimensionality; and Zinbarg, Revelle, Yovel, and Li (2005) compared alpha with several other methods for estimating test score reliability. Each of these papers was published in *Psychometrika*, but other papers have appeared recently in other mainstream methodological journals (e.g., Raykov, 2001; Rodriguez & Maeda, 2006). Also, critical discussions of uses and abuses of alpha have appeared in substantive journals (e.g., Cortina, 1993; Schmitt, 1996).

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

3. Reliability, the Greatest Lower Bound, and Alpha

3.1. Test Score Reliability and the Greatest Lower Bound

The definition of reliability is based on parallel test forms (Novick, 1966; Novick & Lewis, 1967; also, see Lord & Novick, 1968). Let random variable X_j denote the score on item j; for example, $X_j = 0, 1$ for incorrect/correct scoring typical of performance tests, and $X_j = 0, \ldots, m$ for ordered rating scales typical of behavior assessment. The test contains J items. A much-used summary of the item scores is the total score or test score, which is defined as

$$X_+ = \sum_{j=1}^{J} X_j.$$

Let respondents be indexed by i, such that X_{+i} denotes respondent i's total score.

Test score X_{+i} is assumed to suffer from random measurement error. Thus, rather than X_{+i}, one would like to know respondent i's test score without error. This error-free test score is defined operationally, which means, technically and without reference to a situation in real life, as the expectation of X_{+i} across the propensity distribution of independent repetitions of the test to individual i, that is, as $\varepsilon(X_{+i})$ (Lord & Novick, 1968, pp. 29–30). The expectation is better known as the true score, T_i, such that

$$T_i = \varepsilon(X_{+i}).$$

Because the true score is a real number, it could never be the result of adding integer item scores; thus, the + sign does not appear with T. In item response theory (IRT), the propensity distribution appears in the stochastic subject formulation of response behavior at the level of individual items (Holland, 1990).

The difference between a test score resulting from a single test administration and the true score is defined to be the random measurement error,

$$E_i = X_{+i} - T_i,$$

which is a real number and like T does not carry the + sign. Because measurement errors are assumed to originate unpredictably from a random process, they correlate 0 with any other variable unless they are part of that variable (such as X_{+i}). Also, in a population of respondents for which only one test score is available, it is assumed that measurement errors correlate 0 with any other variable unless they are part of that variable.

Parallel tests represent a mathematical definition of independent repetitions of the same test under the same circumstances. Two tests, with test scores X_+ and X'_+, are parallel if

$$(1) \quad T_i = T'_i, \quad \text{for all } i,$$

and, denoting variance by σ^2, if also

$$(2) \quad \sigma^2_{X_+} = \sigma^2_{X'_+}.$$

Thus, (1) an individual has the same long-run test performance on both tests, and (2) the variance of the test scores in the population is the same for both tests. It can easily be shown that these two properties imply that parallel tests have exactly the same psychometric properties. For example, the correlation of X_+ and X'_+ with any other independently measured variable Y is the same for both tests, implying equal validity. The only difference resides in the test scores themselves: that is, in general $X_{+i} \neq X'_{+i}$, which is due to random measurement error.

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

The reliability of the test score X_+ in the population of interest is defined as the product-moment correlation between the scores on X_+ and the scores on a test parallel to this test with scores denoted by X'_+, and the reliability is denoted by $\rho_{X_+X'_+}$. Because one test is parallel to the other, the correlation between the test scores gives the reliability of both X_+ and X'_+ separately. A well-known result is that $0 \leq \rho_{X_+X'_+} \leq 1$. It can further be shown that $\sigma^2_{X_+} = \sigma^2_T + \sigma^2_E$ (and, likewise, $\sigma^2_{X'_+} = \sigma^2_{T'} + \sigma^2_{E'}$), and that consequently for test score X_+,

$$\rho_{X_+X'_+} = \frac{\sigma^2_T}{\sigma^2_{X_+}} = 1 - \frac{\sigma^2_E}{\sigma^2_{X_+}}. \tag{1}$$

Thus, three interchangeable ways of saying that reliability is higher, are that parallel test forms correlate higher, true score variance is greater relative to test score variance, and error variance is smaller relative to test score variance.

Equation (1) shows that the reliability can be estimated if two parallel versions of the test are available or if the true score variance (or, equivalently, the error variance) is available on the basis of one test administration. Because these possibilities are unattainable in practical test research, many alternatives have been proposed (e.g., Guttman, 1945; Nunnally, 1978) that use the data available from a single test administration. The most instructive method is the glb (e.g., Bentler & Woodward, 1980; Jackson & Agunwamba, 1977; Woodhouse & Jackson, 1977). Ten Berge and Sočan (2004) explain the glb as follows. The interitem covariance matrix for observed item scores, $\mathbf{C}_X$, is decomposed into the sum of the interitem covariance matrix for item true scores, $\mathbf{C}_T$, and the interitem error covariance matrix $\mathbf{C}_E : \mathbf{C}_X = \mathbf{C}_T + \mathbf{C}_E$. The interitem error covariance matrix $\mathbf{C}_E$ is diagonal with error variances on the main diagonal and off-diagonal zeroes reflecting that errors correlate zero with any other variable in which they are not included. All three matrices are positive semidefinite (psd; i.e., they do not have negative eigenvalues). The glb problem is solved by finding the nonnegative matrix $\mathbf{C}_E$ for which $\mathbf{C}_T = \mathbf{C}_X - \mathbf{C}_E$ is psd that minimizes

$$r_{X_+X'_+} = 1 - \frac{\text{tr}(\mathbf{C}_E)}{S^2_{X_+}}.$$

This is the glb because it represents the smallest reliability possible given observable covariance matrix $\mathbf{C}_X$ under the restriction that the sum of error variances is maximized for errors that correlate 0 with other variables. Thus, the data obtained from one test administration restrict the real reliability to the interval [*glb*, 1]. This means that when the glb is found to be 0.8, the true reliability has a value in the interval [0.8; 1]. Thus, data from a single test administration restrict the reliability to an interval, whereas data from two parallel tests would yield a point estimate of the reliability. Algorithms for solving the glb problem are discussed by Bentler and Woodward (1980) and Ten Berge, Snijders, and Zegers (1981).

3.2. Definition of Alpha

Let σ^2_j denote the variance of item score X_j and σ_{jk} the covariance between item scores X_j and X_k. Alpha is defined as

$$alpha = \frac{J}{J-1}\left[1 - \frac{\sum_{j=1}^{J}\sigma^2_j}{\sigma^2_{X_+}}\right],$$

or equivalently as

$$alpha = \frac{J}{J-1}\frac{\sum\sum_{j\neq k}\sigma_{jk}}{\sigma^2_{X_+}}. \tag{2}$$

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

This latter form proves to be useful later on. It may be noted that $c \leq alpha \leq 1$, with $c < 0$ if the mean interitem covariance among the J items is negative. This is known to happen sometimes due to accidentally coding both positively and negatively worded personality or attitude items in the same direction.

3.3. Relationship Between Alpha, the glb, and Reliability

Guttman (1945, p. 274) proved that alpha—his λ_3 coefficient—is a lower bound to the reliability, that is, he proved that for J items,

$$alpha \leq \rho_{X_+X'_+}.$$

Novick and Lewis (1967, Theorem 3.1) proved that $alpha = \rho_{X_+X'_+}$ holds if and only if the items in the test are essentially τ-equivalent (τ is sometimes used to denote the true score, i.e., $\tau = T$). Essential τ-equivalence is another mathematical definition of the similarity of different tests (here, items are considered as 1-item tests) that is less restrictive than parallelism. For items j and k, and constant a_{jk}, essential τ-equivalence is defined as

$$T_j = T_k + a_{jk}, \quad \text{for all item pairs } j \neq k.$$

Essential τ-equivalence implies that interitem covariance σ_{jk} is the same for all item pairs ($j \neq k$), and that covariance σ_{jY} is the same for all items ($j = 1, \ldots, J$) and any independently measured variable Y. Like parallelism, essential τ-equivalence is not a realistic condition in test data, so that in real data we have that $alpha < \rho_{X_+X'_+}$.

The glb relates to alpha and the reliability as

$$alpha \leq glb \leq \rho_{X_+X'_+}. \tag{3}$$

Equation (3) is true because $alpha \leq glb$ (Jackson & Agunwamba, 1977), and by definition $glb \leq \rho_{X_+X'_+}$. We know that $alpha = \rho_{X_+X'_+}$ if and only if the items are essentially τ-equivalent. Also, $glb = \rho_{X_+X'_+}$ if the items are essentially τ-equivalent but equality can also be obtained under other conditions. For example (Ten Berge, personal communication), one may use covariance matrix

$$\mathbf{C}_X = \begin{bmatrix} 1 & 1 & 2 \\ 1 & 2 & 3 \\ 2 & 3 & 5 \end{bmatrix} \tag{4}$$

and (7b) from Ten Berge and Sočan (2004) to verify that $glb = 1$. This result implies that $glb = \rho_{X_+X'_+}$ even though the covariances in (4) are unequal, which violates essential τ-equivalence. Thus, it follows that $\mathbf{C}_X = \mathbf{C}_T$ (also, see Ten Berge & Sočan, 2004, (5), first part) and $\mathbf{C}_E = \mathbf{0}$. Also, notice that in this example $alpha = 0.9$.

Equation (3) shows that for real data alpha is not in the interval [glb, 1] of admissible values, and the conclusion can only be that for any observable covariance matrix $\mathbf{C}_X$ alpha provides a value that cannot be a possible value of the reliability based on the knowledge provided by one test administration. One could argue that it does not hurt to use a small lower bound like alpha in practice, because unnecessary low reliability estimates may have the positive effect of stimulating the researcher to do all (s)he can to construct a high-quality test. Much as this is true, following this same line of reasoning accepting an even smaller lower bound such as Guttman's (1945) λ_1 coefficient ($\lambda_1 < alpha$ for finite test length) would even boost that effect. Perhaps it is more reasonable to ask why one would report an estimate of the reliability in the face of much better alternatives, the most prominent one being the glb. Schmitt (1996) warns that lower bounds like alpha may produce gross overestimates of the correlation between test scores when they are corrected for attenuation. The real-data example reported in the next subsection shows that using the glb instead of alpha or another lower bound can make a difference indeed.

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

3.4. Alternatives for Alpha

Borsboom (2006) noted that the degree to which a statistical method is used in empirical research very much depends on its availability in SPSS. Alpha is in SPSS, and so are the other five lower bounds proposed by Guttman (1945). One of them is known under the name of λ_2, and is sometimes reported instead of alpha. Guttman (1945) proved that $alpha \leq \lambda_2$. Alpha and λ_2 are the first two terms of an infinite series of lower bounds in which they are denoted by μ_0 and μ_1 (Ten Berge & Zegers, 1978), respectively. Ten Berge and Zegers (1978) concluded that computing lower bounds from their series beyond λ_2 usually does not produce increases that are worthwhile reporting.

Coefficient λ_2 relates to alpha, the glb, and the reliability as

$$alpha \leq \lambda_2 \leq glb \leq \rho_{X_+X'_+}. \tag{5}$$

In (5), we have that $\lambda_2 \leq glb$ (Jackson & Agunwamba, 1977). Except for the relationship between *glb* and $\rho_{X_+X'_+}$, (5) contains equalities if and only if the items are essentially τ-equivalent. Equality between *glb* and $\rho_{X_+X'_+}$ can also be obtained under different conditions; see (4) for an example. Information about the sampling characteristics of lower bounds is available from several sources (e.g., Feldt, Woodruff, & Salih, 1987). The glb estimate may be biased positively even for samples as large as 1,000 cases, but bias seems to be rather small as the number of items is smaller than 10 (Ten Berge & Sočan, 2004).

The three lower bounds alpha, λ_2, and the glb were computed for a real-data example. The data came from a questionnaire that consists of eight rating scale items, scored 0, 1, 2, 3, by 828 respondents. Each item asked respondents who lived in the vicinity of a malodorous factory how they coped with industrial malodors (Cavalini, 1992). The dimensionality of the data was investigated using principal components analysis (PCA). Researchers typically use PCA to investigate the dimensionality of the data, but better methods may be available to be discussed later on. Alpha, λ_2, and PCA and Varimax rotation were computed by means of SPSS 14.0 (2006), and the glb was computed by means of the program MRFA2.exe (Ten Berge & Kiers, 2003) that can be downloaded from http://www.ppsw.rug.nl/~kiers/.

PCA resulted in eigenvalues for the inter-item correlation matrix $\mathbf{R}_X$ of 3.213, 1.103, and the next six each smaller than 1. The second component showed the contrasts typically suggesting that a rotated 2-factor solution would better explain the correlation structure despite only one eigenvalue being markedly greater than 1. Indeed, Varimax rotation of the first two components resulted in one set of three items with loadings on the first factor greater than 0.7, and another set of four items with loadings on the second factor greater than 0.6 (Table 1). One item had a loading of approximately 0.5 on both factors, but based on content it went better with the items loading highest on the first factor.

For all eight items considered to be in one scale, alpha is 0.007 smaller than λ_2, but λ_2 is 0.067 smaller than the glb (Table 2). The lower bound values for the first of the two 4-item scales resulting from the PCA were nearly as high as those found for the 8-item scale, but the lower bounds for the second 4-item set were smaller by approximately 0.15. Still, λ_2 was 0.074 smaller than the glb in the first scale, and 0.052 smaller than the glb in the second scale. The gap

TABLE 1.
Factor loadings for eight items measuring coping styles.

Factor	Item No.							
	3	5	6	7	9	11	13	14
I	0.17	0.54	0.21	0.75	0.85	0.70	0.15	0.10
II	0.64	0.49	0.60	0.18	0.07	0.24	0.74	0.70

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

TABLE 2.
Lower bounds alpha, λ_2, and the glb, Total Observed Variance (TotObsVar), Total Common Variance (TotComVar), and Explained Common Variance (ECV) for an 8-item scale and two 4-item scales.

	No. Items (J)		
	8	4 (set 1)	4 (set 2)
alpha	0.778	0.736	0.640
λ_2	0.785	0.746	0.644
glb	0.852	0.820	0.696
TotObsVar	8	4	4
TotComVar	4.2294	2.3616	1.6283
ECV	65.39	79.57	85.42

TABLE 3.
Covariance matrix for eight items. Items 5, 7, 9, 11 are in Set 1; and items 3, 6, 13, 14 are in Set 2. Interitem covariance within these sets in bold face.

Item No.	3	5	6	7	9	11	13	14
3	1.110							
5	0.385	0.894						
6	**0.203**	0.259	0.510					
7	0.169	**0.203**	0.122	0.395				
9	0.152	**0.240**	0.096	**0.246**	0.490			
11	0.223	**0.453**	0.177	**0.191**	**0.293**	0.858		
13	**0.264**	0.234	**0.204**	0.141	0.145	0.186	0.622	
14	**0.263**	0.245	**0.140**	0.127	0.124	0.192	**0.287**	0.706

between alpha/λ_2 and the glb was caused by the spread in the interitem covariances (Table 3). This violation of a necessary condition for essential τ-equivalence prevented the three lower bounds from being equal.

The differences reported here are believed to be of practical interest to test constructors and researchers who report a reliability estimate for their test or questionnaire. Moreover, in this data set, there is no convincing reason to report unnecessary small reliability estimates.

4. Alpha as Measure of Internal Consistency

4.1. Drifting Away From Reliability to Internal Consistency

When Cronbach published his classical article in 1951, it was already known that alpha was a lower bound to the reliability, but it is important to realize that at that time several definitions of test score reliability, true score, and random measurement error existed next to one another. The widely accepted foundation of classical test theory as provided later on by Novick (1966) and Lord and Novick (1968) was unknown then. Thus, Cronbach (1951, p. 299) could write: "It has generally been stated that α (i.e., Cronbach's alpha; the author) gives a lower bound to the "true reliability"—whatever that means to that particular writer." As a result, the concept of a lower bound did not seem as compelling to Cronbach as it is nowadays and, instead, much of Cronbach's paper was not about alpha as a lower bound but about analyzing the relationships of alpha with correlations between similar test forms ("similar" is different here from parallel), test-retest correlation, and split-half correlation, and with the factorial composition of the test. This produced several interesting results that were picked up by many psychologists and led to the interpretation of alpha as a measure of the *internal consistency* of a test. It is safe to say

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

that the interpretation of alpha as a measure of internal consistency has gained more foothold in practical test construction and test use than the lower bound interpretation. Before I try to explain this preference, I first ask what internal consistency is.

Schmitt (1996) distinguishes internal consistency from homogeneity, and claims that internal consistency refers to the interrelatedness of a set of items, and homogeneity to the unidimensionality of a set of items. However, this distinction does not convincingly solve terminological confusion. To start with, unidimensionality is not a unitary concept. The concept plays a role both in factor analysis and IRT, and has been defined in different ways. There are similarities, however. Lord and Novick (1968, p. 374, Theorem 16.8.1) proved that if J dichotomous items originate from different dichotomizations of J normal distributions of latent continuous item scores, which have a rank 1 covariance matrix, then the regression of each item on the latent trait is a 2-parameter normal ogive. Takane and De Leeuw (1987) studied the relationship between the factor model and normal ogive IRT models in a more general framework. Independent of factor models, within the class of different unidimensional logistic IRT models such as the 1-, 2-, and 3-parameter models, each model imposes different restrictions on the data, and each model may be seen as representing another definition of unidimensionality. Thus, it seems that in general the concept of unidimensionality is tied to a particular model and in this sense it is clear what unidimensionality means under that particular model.

Internal consistency has not been defined that explicitly, far from it. For example, Cronbach (1951, p. 320) used internal consistency and homogeneity synonymously (cf. Schmitt, 1996), and noted that an internally consistent test is "psychologically interpretable" although this does not mean "that all items be factorially similar." In the jargon of test construction internal consistency often refers to the items "being interrelated" (Schmitt, 1996) but other interpretations are also used regularly. In practical test construction, the use of alpha often goes hand-in-hand with PCA (e.g., Cavalini, 1992; De Hooge, Zeelenberg, & Bruegelmans, 2007). A pervasive albeit informal interpretation of the test's internal consistency is that the first eigenvalue of the interitem correlation matrix is high relative to the second eigenvalue but exactly how high is unclear. This interpretation indeed is different from equating internal consistency with a 1-factor solution or with IRT unidimensionality and leaves open the possibility that different items have varying patterns of factor loadings if more than one factor is retained. This comes close to Cronbach's remark that items need not be factorially similar for the test to be internally consistent. But what this analysis does best is underline the vagueness of the internal consistency concept.

This vagueness has not stopped alpha from becoming a landmark for internal consistency. Remarkably, however, is that a glance at alpha shows that all other things kept equal, its value depends only on the sum of the interitem covariances (2). Thus, all that alpha can reveal about the "interrelatedness of the items" is their *average* degree of "interrelatedness" provided there are no negative covariances, and keeping in mind that alpha also depends on the number of items in the test (Nunnally, 1978, pp. 227–228). Because this says very little if anything about internal consistency no matter how it is defined, one wonders why the internal consistency interpretation of alpha is so persistent. I believe that there are two related reasons.

The first reason is that while several studies have well illuminated the relationships of alpha to other quantities (e.g., Cortina, 1993; Green, Lissitz, & Mulaik, 1977; also see Cronbach, 1988), in particular the factor structure of the test, they have also conveyed the impression that because alpha has something to do with the test's factor structure, its value therefore must express characteristics of this factor structure. This conclusion is logically incorrect and usually not intended by these studies, but it probably has been too compelling to many test constructors and test practitioners to resist. A single number—alpha—that expresses both reliability *and* internal consistency—conceived of as an aspect of *validity* that suggests that items "measure the same thing"—is a blessing for the assessment of test quality. In the meantime, alpha "only" is a lower bound to the reliability and not even a realistic one.

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

The second reason is that after the 1950s, psychometrics has developed to become more mathematically and statistically oriented while psychologists primarily have remained psychologists. One can argue whether psychologists should become better statisticians or whether psychometricians should become better psychologists (Borsboom, 2006), but it is a fact that the two worlds have drifted apart more than anyone should wish. Thus, while much of Cronbach's paper was and still is accessible to many psychologists, the work by Lord, Novick, and Lewis and many others since may have gone unnoticed by most psychologists. This is truly an example of the gap that has grown between psychometrics and psychology and that prevents new and interesting psychometric results, including those that relate alpha to the glb and the test's factor structure, to seep into mainstream psychology.

4.2. Alpha and Internal Test Structure

There is no clear and unambiguous relationship between alpha and the internal structure of a test. This can be demonstrated in a simple way. First, it is shown that a 1-factor test may have any alpha value. Thus, it may be concluded that the value of alpha says very little if anything about unidimensionality. Second, it is shown that different tests of varying factorial composition may have the same alpha value. Thus, it may be concluded that alpha says very little if anything about multiple-factor item structures.

Alpha and Unidimensionality Equal item variances, equal interitem covariances and, consequently, equal interitem correlations are necessary (but not sufficient) for parallel items. Ten Berge and Kiers (1991) advocated the use of minimum rank factor analysis (MRFA) for assessing closeness of the covariance/correlation matrix to unidimensionality. For a 1-factor solution, MRFA determines the diagonal uniquenesses covariance matrix $\mathbf{C}_E$, which produces the smallest sum of the $J-1$ smallest eigenvalues of the difference matrix $\mathbf{C}_X - \mathbf{C}_E$. Thus, the amount of common variance that is left unexplained when the last $J-1$ factors are ignored is minimized, and as a result, the 1-factor solution is the "most-unidimensional" factor solution.

Closeness of the 1-factor solution to unidimensionality is assessed by means of the ratio of the first eigenvalue of $\mathbf{C}_X - \mathbf{C}_E$ and the sum of all J eigenvalues of $\mathbf{C}_X - \mathbf{C}_E$ (Ten Berge & Sočan, 2004). After transforming this ratio to a percentage, the explained *common* variance (ECV) is obtained. Instead of MRFA and the ECV, test constructors often use PCA and the percentage of *observed* variance (POV) corresponding to the first eigenvalue extracted by PCA from the correlation matrix $\mathbf{R}_X$. It should be noted that PCA is based on $\mathbf{C}_E = \mathbf{0}$, and thus provides the "least-unidimensional" factor solution in terms of eigenvalues corresponding to $\mathbf{C}_X - \mathbf{C}_E$.

A 1-factor item structure was operationalized in each of seven tests, each test consisting of six items ($J = 6$) with item variances all equal to $\sigma_j^2 = 0.25$ ($j = 1, \ldots, 6$) and equal positive interitem covariances σ_{jk}. Across tests, from a practical point of view covariances varied from high ($\sigma_{jk} = 0.15$; corresponding to pm-correlation $\rho_{jk} = 0.6$) to low ($\sigma_{jk} = 0.01$; corresponding to pm-correlation $\rho_{jk} = 0.04$). For each of the seven interitem correlation matrices, MRFA was done by means of the program MRFA2.exe (Ten Berge & Kiers, 2003), and PCA was done by means of SPSS 14.0 (2006) syntax code. MRFA.2.exe also produces the glb. Alpha was computed by means of SPSS 14.0 syntax code, and was compared to the glb.

ECV is 100% for all seven interitem correlation matrices but POV, which is the quantity that most test constructors use for assessing unidimensionality, starts at 66.67% and then drops gradually to 20% (Table 4). Thus, ECV indicates perfect unidimensionality whereas POV suggests factor solutions that move away from unidimensionality. This conclusion is amplified if one also takes the eigenvalues of correlation matrix $\mathbf{R}_X$ (PCA) into consideration. Many researchers probably would take the last two or three sets of eigenvalues (Table 4) as evidence of multidimensionality (i.e., the items each correspond to unique factors) instead of unidimensionality. Because

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

TABLE 4.
Eigenvalues (EV) of observable correlation matrix $\mathbf{R}_X$, percentage of observed variance (POV) explained by the first principal component, ECV, alpha, and glb, for tests with $J = 6$, $\sigma_j^2 = 0.25$ ($j = 1, \ldots, 6$), and σ_{jk}/ρ_{jk} constant per test and variable across tests.

		σ_{jk}/ρ_{jk}						
		0.15/0.60	0.12/0.48	0.09/0.36	0.06/0.24	0.03/0.12	0.02/0.08	0.01/0.04
No. EV	1	4.00	3.40	2.80	2.20	1.60	1.40	1.20
	2	0.40	0.52	0.64	0.76	0.88	0.92	0.96
	3	0.40	0.52	0.64	0.76	0.88	0.92	0.96
	4	0.40	0.52	0.64	0.76	0.88	0.92	0.96
	5	0.40	0.52	0.64	0.76	0.88	0.92	0.96
	6	0.40	0.52	0.64	0.76	0.88	0.92	0.96
POV		66.67	56.67	46.67	36.67	26.67	23.33	20.00
ECV		100	100	100	100	100	100	100
alpha		0.90	0.85	0.77	0.65	0.45	0.34	0.20
glb		0.90	0.85	0.77	0.65	0.45	0.34	0.20

the covariance matrices are typical of essential τ-equivalence, it follows that *alpha* = *glb*. Table 4 also shows that as inter-item covariance drops while keeping everything else constant, alpha and the glb drop from 0.90 to 0.20, that is, from high to low.

The seven examples in Table 4 each represent cases of unidimensionality: From left to right, the signal becomes weaker while the noise (due to unique factors and measurement error) becomes stronger. But all the time there is one signal—unidimensionality—correctly identified by ECV = 100. The reliability quantifies the degree to which test scores can be repeated under the same circumstances. As the signal in the data becomes weaker, alpha and the glb become smaller, as they should.

Alpha and Multidimensionality Multidimensionality was operationalized by means of three tests, again each consisting of six items ($J = 6$), with item variances equal to $\sigma_j^2 = 0.25$ ($j = 1, \ldots, 6$), and interitem covariances σ_{jk} such that: (1) they were positive and equal within clusters of items; (2) they were zero between items from different clusters; and (3) the sum of all $J(J-1)$ covariances was constant across different matrices $\mathbf{C}_X$. Condition 3 implies the same alpha for each covariance matrix. Table 5 shows the lower triangles of the covariance matrices $\mathbf{C}_X$ with three 2-item clusters, two 3-item clusters, and one 6-item cluster, respectively.

The first two sets of eigenvalues from $\mathbf{R}_X$ each suggest the correct dimensionality of the tests, while the ECV shows that $\mathbf{R}_X$ is remote from unidimensionality. The third set of eigenvalues would probably lead several researchers to conclude that there is one common albeit weak factor, but the ECV suggests perfect unidimensionality. Coefficient alpha equals 0.533 for all three covariance matrices, irrespective of dimensionality. Interestingly, the glb is highest for the 3-factor case and lowest for the 1-factor case (in the latter case, the glb coincides with alpha because $\mathbf{C}_X$ satisfies a necessary condition for essential τ-equivalence; also, see Table 4). More important, alpha does not provide information on the internal structure of the test as it is so often claimed.

Going back to the real-data example discussed previously, it is interesting to see (Table 2) that ECV for the 8-item scale suggests that the scale is remote from unidimensionality. Both 4-item scales have high ECV values suggesting near-unidimensionality, but once more it is clear that unidimensionality or lack thereof has nothing to do with reliability.

Moreover, alpha depends on the number of items J, and our examples can be adapted simply to show that alpha grows as J grows (Cortina, 1993; Green et al., 1977). For example, for $J = 12$, $\sigma_j^2 = 0.25$ ($j = 1, \ldots, 12$), and covariance structures with three 4-item clusters ($\sigma_{jk} = 0.20$

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

TABLE 5.
Covariance matrices $\mathbf{C}_X$, EVs based on corresponding correlation matrix $\mathbf{R}_X$, ECV, glb, and alpha.

$\mathbf{C}_X$						EV $\mathbf{R}_X$	ECV	glb	alpha
0.25						1.8	33.33	0.889	0.533
0.20	0.25					1.8			
0.00	0.00	0.25				1.8			
0.00	0.00	0.20	0.25			0.2			
0.00	0.00	0.00	0.00	0.25		0.2			
0.00	0.00	0.00	0.00	0.20	0.25	0.2			
0.25						1.8	50.00	0.667	0.533
0.10	0.25					1.8			
0.10	0.10	0.25				0.6			
0.00	0.00	0.00	0.25			0.6			
0.00	0.00	0.00	0.10	0.25		0.6			
0.00	0.00	0.00	0.10	0.10	0.25	0.6			
0.25						1.80	100.00	0.533	0.533
0.04	0.25					0.84			
0.04	0.04	0.25				0.84			
0.04	0.04	0.04	0.25			0.84			
0.04	0.04	0.04	0.04	0.25		0.84			
0.04	0.04	0.04	0.04	0.04	0.25	0.84			

within clusters), two 6-item clusters ($\sigma_{jk} = 0.12$) and one 12-item cluster ($\sigma_{jk} = 0.0545454$) such that each time $\sum\sum_{j \neq k} \sigma_{jk} = 7.2$, $alpha = 0.770$ in all three cases.

5. Is There a Future for Alpha?

Lord and Novick (1968) discussed reliability as repeatability of individual test performance described by the individual's propensity distribution. The propensity distribution shows the influence of random measurement error across an infinite number of parallel test administrations. However, due to the practical impossibility to administer the same test to the same individuals repeatedly—even twice is nearly impossible—one has to resort to a random sample of individuals who have been administered the test once, and then estimate the reliability on the basis of this single administration. The glb shows that such data limit the range of possible reliability values to [*glb*, 1] but also that a perfect reliability cannot be ruled out one the basis of one test administration. An interesting question is whether single-administration test data can provide information about individuals' propensity distributions at all.

Molenaar (2004; also, see Borsboom, 2005, pp. 68–81) noted that in general a single-administration sample of test scores does not contain information about the individuals' propensity distributions unless both types of distributions—between individuals as in single-administration data and within individuals as in propensity distributions—obey restrictive distributional properties. He contended that most psychological phenomena do not agree with these assumptions. Other authors also noticed that statements about individuals are problematic when only single-administration data are available. For example, Ellis and Van den Wollenberg (1993) showed that IRT models do not hold for individuals unless the assumption of local homogeneity is added to the models. Molenaar (2004) reported that a (Big) 5-factor personality structure that was found at the group level on the basis of a sample of observations collected at one point in time did not correspond to the different factorial structures characteristic of different individuals

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

who were repeatedly tested by means of the same personality inventory (Molenaar, 2004). This result seems to have relations to the phenomenon that particular individuals are insensitive to certain personality traits, which has become known as lack of traitedness (Tellegen, 1988). Lack of traitedness may be the cause of atypical patterns of scores on items from personality inventories (Reise & Waller, 1993).

Likewise, there is no reason whatsoever to assume that the propensity distributions of different persons must be identical to one another and to the between-persons distribution based on single-administration data. This means that single-administration test data may contain little or no information about propensity distributions. The use of the standard measurement error,

$$\sigma_E = \sigma_X \sqrt{1 - \rho_{X_+ X'_+}},$$

in the practice of psychological testing was born out of this inherent limitation of single-administration test data. The application of the standard measurement error assumes that each individual was tested with the same accuracy but classical test theory does not make this assumption nor is there much reason to expect a priori that people would produce the same propensity distributions when given the opportunity. Indeed, Lord (1960) studied distributions of measurement errors that varied across the true score level, and IRT uses the Fisher information function to estimate a standard error dependent on the scale of measurement. Such improvements recognize the improbability of the same accuracy of measurement for every tested individual but cannot be considered realistic as long as their assumptions have not been put to the test in real data. That is, one needs to study real propensity distributions to find out how standard errors are related to the scale of measurement, and until then the results provided by Lord and IRT are properties of statistical models, not of real behavior.

The problem with discussions like this one is that while (I believe) they make a good point, the practical test user needs to make decisions about the treatment of individual clients or patients and cannot afford to sit back and wait until science comes up with the final solution. Thus, it seems best to end with a number of conclusions about alpha and reliability, and find out what is the next best thing for alpha and reliability in the absence of available propensity distributions.

6. Conclusions

On the basis of the previous discussion, the following five conclusions seem to be in order:

1. In practice, alpha attains values that are outside the range of possible values of the reliability that can be derived from a single test administration. Comparing alpha with the glb gives an impression of the degree to which alpha is wrong. The difference can easily be tenths depending on the exact properties of the test under consideration.
2. Many lower bounds exist between alpha and the glb, and the lower bounds proposed by Guttman (1945) are all in SPSS thus eliminating the "not in SPSS" argument often heard in practice. It is difficult to defend convincingly using one of the smallest lower bounds, alpha, given the availability of many *greater* lower bounds and the glb. The only reason to report alpha is that top journals tend to accept articles that use statistical methods that have been around for a long time such as alpha. Reporting alpha in addition to a greater lower bound may be a good strategy to introduce and promote a better reliability estimation practice.
3. The best lower bound and the only one attaining a realistic value, however, is the glb. The glb is available from several sources and easy to obtain (Ten Berge & Sočan, 2004). Because the glb can be seriously positively biased for lower reliability values, samples smaller than, say, 1,000 cases, and test lengths exceeding, say, 10 items, more work on bias correction is badly

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

needed (e.g., Shapiro & Ten Berge, 2000; Verhelst, 1998) and psychometrics might spend more energy in favor of this just cause. Once a good bias correction is found, one cannot get around the glb anymore to replace alpha (and all other lower bounds).

4. Alpha is not a measure of internal consistency. Neither is it a measure of the degree of unidimensionality (also, see Ten Berge & Sočan, 2004). Alpha has been shown to correlate with many other statistics and much as these results are interesting, they are also confusing in the sense that without additional information, both very low and very high alpha values can go either with unidimensionality or multidimensionality of the data. But given that one needs the additional information to know what alpha stands for, alpha itself cannot be interpreted as a measure of internal consistency.
5. Statistical results based on a single test administration convey little if any information about individuals' measurement accuracy reflected by their propensity distributions. This does not seem to be an insurmountable problem when a test is used for comparing mean scores between different groups or correlations between variables in a nomological network, but even then one has to be aware that "averaging out" the individual causes the means and correlations to lose their psychological meaning (Borsboom, 2005). For drawing conclusions about individuals on the basis of test scores, the best one can do is to use tests that consist of many items and have a reliability—be it estimated by Cronbach's alpha—that pushes 1. More generally, it is recommended to use as much information about the individual as possible (e.g., Emons, Sijtsma, & Meijer, 2007).

Acknowledgements

I am grateful to Wilco H. M. Emons, Brian W. Junker, Roger E. Millsap, Jos M. F. ten Berge, and L. Andries van der Ark for their critical comments to an earlier draft of this paper. Of course, the views presented here are the author's responsibility.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References

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

Borsboom, D. (2005). *Measuring the mind. Conceptual issues in contemporary psychometrics*. Cambridge: Cambridge University Press.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika*, *71*, 425–440.

Campbell, D. T. (1960). Recommendations for APA tests regarding construct, trait or discriminant validity. *American Psychologist*, *15*, 546–553.

Cavalini, P. M. (1992). *It's an ill wind that brings no good. Studies on odour annoyance and the dispersion of odorant concentrations from industries*. Ph.D. thesis, University of Groningen, The Netherlands.

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*, 63–70.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, *78*, 98–104.

De Hooge, I. E., Zeelenberg, M., & Breugelmans, S. M. (2007). Moral sentiments and cooperation: Differential influences of shame and guilt. *Cognition and Emotion*, *21*, 1025–1042.

Ellis, J. L., & Van den Wollenberg, A. L. (1993). Local homogeneity in latent trait models. A characterization of the homogeneous monotone IRT model. *Psychometrika*, *58*, 417–429.

Emons, W. H. M., Sijtsma, K., & Meijer, R. R. (2007). On the consistency of individual classification using short scales. *Psychological Methods*, *12*, 105–120.

Feldt, L. S., Woodruff, D. J., & Salih, F. A. (1987). Statistical inference for coefficient alpha. *Applied Psychological Measurement*, *11*, 93–103.

Green, S. B., Lissitz, R. W., & Mulaik, S. A. (1977). Limitations of coefficient alpha as an index of test unidimensionality. *Educational and Psychological Measurement*, *37*, 827–838.

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika*, *10*, 255–282.
Hayashi, K., & Kamata, A. (2005). A note on the estimator of the alpha coefficient for standardized variables under normality. *Psychometrika*, *70*, 579–586.
Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika*, *55*, 577–601.
Hoyt, C. (1941). Test reliability estimated by analysis of variance. *Psychometrika*, *6*, 153–160.
Jackson, P. H., & Agunwamba, C. 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*, 567–578.
Kistner, E. O., & Muller, K. E. (2004). Exact distributions of intraclass correlation and Cronbach's alpha with Gaussian data and general covariance. *Psychometrika*, *69*, 459–474.
Kuder, G. F., & Richardson, M. W. (1937). The theory of estimation of test reliability. *Psychometrika*, *2*, 151–160.
Lord, F. M. (1960). An empirical study of the normality and independence of errors of measurement in test scores. *Psychometrika*, *25*, 91–104.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Wesley.
Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology—This time forever. *Measurement*, *2*, 201–218.
Novick, M. R. (1966). The axioms and principal results of classical test theory. *Journal of Mathematical Psychology*, *3*, 1–18.
Novick, M. R., & Lewis, C. (1967). Coefficient alpha and the reliability of composite measurements. *Psychometrika*, *32*, 1–13.
Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.
Raykov, T. (2001). Bias of coefficient alpha for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.
Reise, S. P., & Waller, N. G. (1993). Traitedness and the assessment of response pattern scalability. *Journal of Personality and Social Psychology*, *65*, 143–151.
Rodriguez, M. C., & Maeda, Y. (2006). Meta-Analysis of coefficient alpha. *Psychological Methods*, *11*, 306–322.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment*, *8*, 350–353.
Shapiro, A., & Ten Berge, J. M. F. (2000). The asymptotic bias of minimum trace factor analysis, with applications to the greatest lower bound to reliability. *Psychometrika*, *65*, 413–425.
SPSS Inc. (2006). *SPSS 14.0 for Windows* (computer software). Chicago: Author.
Takane, Y., & De Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393–408.
Tellegen, A. (1988). The analysis of consistency in personality assessment. *Journal of Personality*, *56*, 621–663.
Ten Berge, J. M. F., & Kiers, H. A. L. (1991). A numerical approach to the exact and the approximate minimum rank of a covariance matrix. *Psychometrika*, *56*, 309–315.
Ten Berge, J. M. F., & Kiers, H. A. L. (2003). *The minimum rank factor analysis program MRFA* (Internal report). Department of Psychology, University of Groningen, The Netherlands.
Ten Berge, J. M. F., & Sočan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika*, *69*, 613–625.
Ten Berge, J. M. F., & Zegers, F. E. (1978). A series of lower bounds to the reliability of a test. *Psychometrika*, *43*, 575–579.
Ten Berge, J.M.F., Snijders, T.A.B., & Zegers, F.E. (1981). Computational aspects of the greatest lower bound to the reliability and constrained minimum trace factor analysis. *Psychometrika*, *46*, 201–213.
Van Zyl, J. M., Neudecker, H., & Nel, D. G. (2000). On the distribution of the maximum likelihood estimator of Cronbach's alpha. *Psychometrika*, *65*, 271–280.
Verhelst, N. D. (1998). *Estimating the reliability of a test from a single test administration* (Measurement and Research Department Report 98-2). Arnhem, The Netherlands, CITO National Institute for Educational Measurement.
Watson, J. D., & Crick, F. H. C. (1953). Molecular structure of nuclied acids—a structure for deoxyribose nucleid acid. *Nature*, *171*, 737–738.
Woodhouse, B., & Jackson, P. H. (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*, 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*, 123–133.

Published Online Date: 11 DEC 2008

Downloaded from https://www.cambridge.org/core. 09 Jan 2025 at 13:18:16, subject to the Cambridge Core terms of use.