PSYCHOMETRIKA—VOL. 74, NO. 1, 169–173
MARCH 2009
DOI: 10.1007/S11336-008-9103-Y

RELIABILITY BEYOND THEORY AND INTO PRACTICE

KLAAS SIJTSMA

TILBURG UNIVERSITY

The critical reactions of Bentler (2009, doi:10.1007/s11336-008-9100-1), Green and Yang (2009a, doi:10.1007/s11336-008-9098-4; 2009b, doi:10.1007/s11336-008-9099-3), and Revelle and Zinbarg (2009, doi:10.1007/s11336-008-9102-z) to Sijtsma's (2009, doi:10.1007/s11336-008-9101-0) paper on Cronbach's alpha are addressed. The dissemination of psychometric knowledge among substantive researchers is discussed.

Key words: common factor reliability, Cronbach's alpha, structural equation modeling of reliability.

Bentler (2009), Green and Yang (2009a), and Revelle and Zinbarg (2009) provided useful and enlightening reactions to my critical paper on Cronbach's (1951) alpha (Sijtsma, 2009), and Green and Yang (2009b) proposed a new method for reliability estimation. The reactions show much of the wealthy harvest a century of reliability theory has brought onto psychology. Nowadays, several substantive researchers use methods alternative to alpha for estimating reliability, but they represent a minority. One of the challenges the psychometric community is facing today is finding ways to disseminate the alternative and usually better reliability methods among psychological researchers. In this respect, Revelle and Zinbarg are right that an obsolete black MS-dos screen, as the one I used for estimating the greatest lower bound (glb) to the reliability, is unlikely to divert the mainstream of psychological researchers from alpha to the glb. I address the topic of dissemination of psychometric knowledge, including superior methods for reliability estimation, after I have provided a brief rejoinder to the discussants.

1. Rejoinder

The purposes of my paper were to recall that better reliability estimates exist than alpha, to warn against the misinterpretation of alpha as an index of internal consistency, and to suggest that reliability is perhaps not that important when it comes to assessing the accuracy of diagnosing individuals (e.g., see Cronbach, 2004). Perhaps my reticence with respect to the third topic has kept the discussants from giving the issue attention. They seem to agree with my rejection of the internal consistency interpretation of alpha, but keep using the misleading terminology in their contributions. Since alpha does not index whether the test is unidimensional, 1-factorial, consistent, or homogeneous, why do we not get rid of the term of internal consistency altogether?

Most of the discussion concentrates on alternatives for alpha and, more generally, true-score reliability. Green and Yang (2009b) actually present a new method for estimating reliability based on nonlinear structural equation modeling (SEM). This new method corrects for the discreteness of item scores, which might distort reliability estimates obtained from a linear model that incorrectly assumes continuity. More generally, the four reactions have in common that they take the confirmatory factor model estimated in the SEM context as point of departure. This suggests that present-day theorizing about reliability is heading toward a latent-variable modeling approach.

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

Classical test theory (CTT) assumes random measurement errors that correlate zero with any other variable in which they are not included, and true scores that include everything systematic in the test score. CTT assesses test-score reliability as the proportion of true-score variance, without considering the composition of the true score. SEM reliability, however, is concerned with the composition of the true score. SEM approaches to reliability decompose the true-score variance into different variance components, and the researcher has to decide which variance components contribute to test-score reliability (Bentler, 2009). Examples are variance components due to auxiliary attributes that are measured next to the dominant attribute of interest, and variance components due to unwanted influences on item performance, such as response styles.

1.1. Correlated Errors and Violation of Essential τ-Equivalence

Bentler (2009) discusses alternatives to alpha based on a decomposition of the test score into a common part, a specific part, and a random error part. He defines what he calls internal-consistency reliability as the proportion of common variance rather than true-score variance. Obviously, the former cannot exceed the latter, and alpha is shown not to exceed the proportion of common variance. Bentler argues that correlated errors, which are the result of auxiliary unwanted influences on the performance on several items producing an unwanted systematic variance component, may cause alpha to overestimate reliability as the proportion of common variance.

Green and Yang (2009a) discuss how violations of the CTT assumption of uncorrelated errors produce bias in alpha. Correlated errors are assumed to arise from the nesting of subgroups of items within, for example, different reading texts; order effects due to prior item performance affecting performance on later items; response sets due to a particular wording of subsets of items; and transient errors due to respondents' feelings and attitudes that are active during one test taking but not during another. Like Bentler (2009), Green and Yang (2009a) conclude that the effect on alpha often is inflation to the point of an overestimation of reliability (also, see Raykov, 2001, for a numerical example), but they also notice that alpha may be too low. They point out that a complicated design is needed if one were to disentangle in a confirmatory factor analysis framework variance due to correlated errors involving at least two items from variance due to item-specific factors.

All of this is important in the SEM framework but none of it matters when one is only interested in knowing to which degree test scores would be different upon repetition of the measurement procedure. The CTT angle to reliability thus ignores the composition of the true score, interestingly allocating this topic to the validity study of the test.

For alpha to be equal to the reliability, the items have to be essential τ-equivalent. Green and Yang (2009a) argue rightly that essential τ-equivalence requires unidimensionality of the item set but that in practice the assumption likely is violated because most tests measure several narrow group factors in addition to a general factor. Even purely unidimensional tests are expected to have items that differ with respect to true score variance, thus violating essential τ-equivalence. Hence, Cronbach's alpha underestimates true reliability. These authors advocate confirmatory factor analysis for checking whether the assumptions of uncorrelated errors and essential τ-equivalence hold in the data and discuss SEM for estimating reliability. For example, SEM may be used for testing whether the items are essential τ-equivalent and simultaneously estimating the reliability of the unweighted sum score based on the items (Raykov, 1997). A weakness of SEM is that a good reliability estimate requires a fitting model (Green & Yang, 2009a), just as the equality of alpha to the reliability requires the items to be essential τ-equivalent.

1.2. Estimating Reliability by Means of Structural Equation Modeling

SEM is highly flexible in that group factors and multidimensional factor structures can be modeled, and the test-score reliability taking such factor structures into account simultaneously

estimated. For example, Raykov and Shrout (2002) discuss a test which consists of different subsets of items, each subset measuring another factor, and for this heterogeneous test these authors demonstrate that alpha is considerably smaller than the reliability estimated from the item loadings of the multi-factor structure.

A question that occurred to me while reviewing this literature was whether it is a good idea to model multidimensionality for the purpose of estimating test-score reliability, and not work the other way around. For example, in an exploratory data analysis SEM could be used to uncover the factorial structure of the item set as part of the validity study of the test. When items are found to measure something else than the attribute of interest they could be removed from the test, and reliability estimated for the test score on the remaining items. The more usual confirmatory context of SEM applications would entail that a hypothesized factorial item structure is tested that is based on a substantive theory about the attribute of interest. When the hypothesis is supported, this might entail that dimensionally distinct item subsets should be distinguished. These subsets could then be considered separate subtests, and reliability estimated for the test scores on each of the subtests separately. Green and Yang (2009a) note that purely unidimensional tests do not exist in practice, and that good construct coverage requires small deviations from unidimensionality.

The strategy outlined avoids the possible confounding of issues of reliability and validity, which could result from the structural modeling of multi-factor data with the aim of estimating test-score reliability. SEM seems to be best conceived of as a powerful tool for understanding or testing the test composition, but also for detecting unwanted influences resulting in correlated errors. Green and Yang (2009a) notice the danger of a similar conceptual blur when discussing McDonald's (1999) ω_h coefficient, which represents only the proportion of variance attributable to the general factor in the data, but Revelle and Zinbarg point out that ω_h should be used as an index of common-factor variance among the items.

1.3. Bias in Reliability Estimates

When data are nearly unidimensional, SEM may be used for estimating test-score reliability but the glb would also be a good candidate. I mentioned in my contribution (Sijtsma, 2009) that the glb estimate may be positively biased, and Bentler explains that this tends to happen in samples smaller than $N = 1{,}000$ because of the high degree of capitalization on sampling fluctuation in the covariance matrix, a point also alluded to by Green and Yang (2009a).

One could ask whether this is a typical glb problem. Raykov (1997) warns that SEM reliability estimates may be misleading with small samples, and thus advises using large samples. Ten Berge and Sočan (2004) argue that the problem also happens with other reliability estimates, which are similar in magnitude to the glb. One of these reliability estimates is McDonald's ω_t coefficient, discussed and advocated by Revelle and Zinbarg, which estimates the sum of the item error variances from the sum of the unique item variances, as one minus the items' communalities. Revelle and Zinbarg show that in many example data sets McDonald's ω_t coefficient exceeds the glb, and advise to use the ω_t coefficient.

Thus, it appears that not only the glb, but also SEM reliability and McDonald's ω_t suffer from bias problems. Having reached this conclusion, perhaps a comparative study of the three methods should be advised.

2. Dissemination of Knowledge

Historically, psychological issues have been the driving force behind the development of psychometric methods, beginning most convincingly with the work of Spearman on intelligence,

factor analysis, and test-score reliability, and continued by Thurstone, Cronbach, Guilford, and many others. As psychometrics developed into a more mature area, psychometricians began looking for new topics, and these were found in statistics and computer science perhaps more than in psychology. This not only weakened the connection between psychological impetus and psychometric method but also created a psychometrics that was mathematically more demanding for psychologists. The result of this loosened tie in combination with more demands caused many new psychometric tools to go unnoticed in psychology.

Reliability theory is a good example of this development. Having originated in psychology, parallel-test reliability, split-half reliability, and lower-bound reliability became standard tools in every psychologist's tool kit. The glb, (nonlinear) SEM reliability, and McDonald's ω_t were developed much later, required a deep knowledge of matrices and statistics, and remained unknown among researchers. Bentler (2009) suggests using the EQS package for computing the glb, Green and Yang (2009b) offer SAS code for nonlinear SEM reliability, and Revelle and Zinbarg (2009) suggest one uses the R software they offer on their website. These are excellent suggestions, but like the coefficients they compute EQS, SAS, and R may not belong to the researcher's toolkit. Even the availability of these methods in SPSS (Borsboom, 2006) would not guarantee their regular use, but it would certainly not harm anyone if they were included.

Many psychometricians think hard about statistical problems that they derive from substantive research and do their best to device user-friendly software and many psychologists use SEM and other advanced methods. But the incorporation of novel methods in data analysis goes at a slow pace. Psychometrics should continue developing relevant tools and easy-to-use software, and provide support with data analysis, and perhaps even more than they have done so far. Interestingly, Cronbach's (1954) Psychometric Society presidential address could serve as a beacon. Psychometric tools should solve recognizable problems and software should be as user-friendly as SPSS, that is, if we value that our tools have impact on substantive research and do not get lost in the no-man's-land between theory and practice.

Acknowledgements

I am grateful to the discussants for providing their reactions to my paper on Cronbach's alpha, and to Wilco H.M. Emons, Jos M.F. ten Berge and Ruud J.H. van Keulen for discussions on issues of interest to this rejoinder. All views expressed are mine, however.

References

Bentler, P.A. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*. doi:10.1007/s11336-008-9100-1.

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

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

Cronbach, L.J. (1954). Report on a psychometric mission to Clinicia. *Psychometrika*, *19*, 263–270.

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

Green, S.A., & Yang, Y. (2009a). Commentary on coefficient alpha: a cautionary tale. *Psychometrika*. doi:10.1007/s11336-008-9098-4.

Green, S.A., & Yang, Y. (2009b). Reliability of summed item scores using structural equation modeling: an alternative to coefficient alpha. *Psychometrika*. doi:10.1007/s11336-008-9099-3.

McDonald, R.P. (1999). *Test theory: a unified approach*. Mahwah: Erlbaum.

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement*, *21*, 173–184.

Raykov, T. (2001). Bias of coefficient α for fixed congeneric measures with correlated errors. *Applied Psychological Measurement*, *25*, 69–76.

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*, 195–212.

Revelle, W., & Zinbarg, R.E. (2009). Coefficients alpha, beta, omega and the glb: comments on Sijtsma. *Psychometrika*. doi:10.1007/s11336-008-9102-z.

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*. doi:10.1007/s11336-008-9101-0.

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.

Manuscript Received: 12 NOV 2008
Published Online Date: 23 DEC 2008