Developments in Survey Methodology
Anuška Ferligoj and Andrej Mrvar (Editors)
Metodološki zvezki, 15, Ljubljana: FDV, 2000

Assessment of Reliability when Test Items are not Essentially τ-Equivalent

Gregor Sočan[1]

Abstract

Estimation of reliability has been a major issue in the 20th century psychometrics; so it is surprising that in practice reliability analysis is usually limited to the computation of α and retest coefficients. Namely, it is well-known that coefficient α is an accurate measure of reliability only if the test items are essentially τ-equivalent; in other cases, it is a lower bound for reliability. In the present paper, some alternative methods which do not require so strict assumptions are described. Probably the most interesting among them are Jöreskog's ML analysis of congeneric measures and Jackson and Agunwamba's greatest lower bound for reliability. These methods' strengths and weaknesses and possibilities for use in psychometric practice are critically discussed. The procedures and their properties are illustrated on several sets of simulated and real (Big Five Questionnaire standardisation, national final high-school examination) data sets. The results show how the adoption of an incorrect measurement model can cause severe underestimation of the reliability coefficient.

1 Introduction

Assessment of reliability and especially reliability as internal consistency has certainly been among the most central issues in psychometrics during this century. Nevertheless, it seems that much of the reliability theory is ignored in psychometric practice: most practitioners use only coefficient α as the internal consistency measure. Unfortunately, they often forget that the use of coefficient α as a reliability estimate rests on certain assumptions which are hardly ever completely met.

[1] Department of psychology, University of Ljubljana, Aškerčeva 2, SI-1001 Ljubljana, Slovenia.

I am indebted to Valentin Bucik, Gaj Vidmar and two anonymous reviewers for giving valuable comments to the manuscript, to Valentin Bucik for providing me the BFQ data and to Henk Kiers and Jos ten Berge for providing me a copy of MRFA2 program.

In the present paper, three different approaches to the assessment of reliability as internal consistency are compared, each of them based on variously rigorous assumptions. All of them follow from the classical test theory.

1.1 General formula for reliability and coefficient α

The core of the classical reliability theory (as elaborated in, e.g., Lord and Novick, 1974) is the additive decomposition of observed scores into true and error scores. True score is defined as the expected value of observed score. Errors are by definition uncorrelated with true scores and supposed to be uncorrelated with other errors. It follows that variance of observed scores is the sum of true and error variance; reliability is then defined as the ratio of true and observed variance.

Considering these basic statements, one can construct the following expression, which is called here the general formula for reliability of a linear composite:

$$r_{SS} = 1 - \frac{\sum \sigma_i^2 - \sum r_{ii}\sigma_i^2}{\sigma_S^2} \tag{1}$$

where S denotes the sum of component variables, which can be items, tests or any other experimentally independent measurements; i denotes the i-th component variable, σ_i is standard deviation of the variable i and r_{ii} stands for reliability of the variable i (for a somewhat different interpretation of the formula see Nunnally and Bernstein, 1994). For brevity, the component variables are termed “items” in this paper. To use eq. (1) one has to know reliabilities of the items. It is obvious that one cannot determine true and error variances by one measurement only — it is necessary to perform several measurements, and at that point one has to make certain assumptions about relationships between true scores on these measurements. One possibility is to take the average covariance of item i with all other items divided by variance of item i as a reliability estimate. It can be shown that in this case eq. (1) can be expressed in the form known as Cronbach’s coefficient α:

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum \sigma_i^2}{\sigma_S^2}\right] \tag{2}$$

It is well known that α is a lower bound for reliability. Usually it is not the best lower bound, but it has some strong advantages over its alternatives:

- its computation is relatively simple, since it requires only the computation of item and total variances;
- the sampling distribution is known, so it is possible to determine confidence intervals for the population coefficient α (Woodward and Bentler, 1978; Feldt, Woodruff and Salih, 1987);
- its computation does not capitalise on chance, and therefore sample estimates of α are usually only negligibly biased (see Feldt, Woodruff, and Salih, 1987);
- it is easy to comprehend.

These advantages made α by far the most popular measure of internal consistency. Unfortunately, its users often forget that α equals reliability only when the items are essentially τ-equivalent, that is, when true scores are perfectly correlated and have equal variances (Lord and Novick, 1974). It follows that all covariances of observed scores are equal, too. If items have different true variances and if they are not unidimensional (which means that true scores are not perfectly correlated), then α will underestimate true reliability. Raykov (1997a) developed an algebraic approach to determination of the size of this underestimation, but it requires use of structural equation modelling. Unfortunately, in practice α will almost always underestimate true reliability. For example, many psychological tests have dichotomously scored items. Since items have different difficulties so that the test can discriminate well over the whole range of scores, the variances of items are necessarily different. If reliabilities are not inversely related to observed variances, true variances will be different, too. Also, factor analyses of presumably unidimensional psychological scales (especially in the personality domain) almost always show that these scales are multidimensional at least to certain extent, which prevents items' true scores to be perfectly intercorrelated.

In most cases, researchers are not very concerned about this problem, since underestimation of reliability is considered to be conservative and therefore not very dangerous. But in some cases it can lead to excessively liberal conclusions. For example, if one performs the correction for attenuation, underestimated reliability will result in *overestimated* corrected correlation (for details see Lord and Novick, 1974: 138).

1.2 Jöreskog's analysis of congeneric measures (ACM)

The analysis of congeneric measures, as developed by Jöreskog (1971), can serve as an alternative to coefficient α if items are suspected to have different true variances. For example, if test items are not scored on the same scale, the scale differences will cause differences between observed variances. If the items are approximately equally reliable, true variances will not be the same, either.

Congeneric measures have pairwise perfectly correlated true scores, but may have different true variances. The rationale of Jöreskog's approach is as follows. If all true scores are pairwise perfectly correlated, then there exists a latent variable τ which is linearly related to all true scores:

$$T_i = a_{i\tau} + b_{i\tau}\tau \tag{3}$$

Now, we can express an observed score as the sum of the true score and error:

$$X_i = a_{i\tau} + b_{i\tau}\tau + E_i \tag{4}$$

If we scale τ to zero mean and unit variance, it follows that

$$\sigma^2_{X_i} = b^2_{i\tau} + \sigma^2_{E_i} \tag{5}$$

Therefore, reliability is equal to the square of the slope divided by the total variance:

$$r_{ii} = \frac{b^2_{i\tau}}{\sigma^2_{X_i}} = \frac{b^2_{i\tau}}{b^2_{i\tau} + \sigma^2_{E_i}} \tag{6}$$

It is not possible to solve for $b_{i\tau}$ analytically: an iterative numerical method is needed to obtain parameter estimates. Jöreskog proposed using maximum likelihood estimation. He also showed that this problem is a case of one factor confirmatory factor-analytic model.

Jöreskog's suggestion was to use whole tests as variables, but the variables can well be items. In that case equation (6) represents reliability of a single item, while it seems evident to use equation (1) to estimate the reliability of the total test, although Jöreskog did not explicitly mention this possibility. Somewhat more technically complex approach was proposed by Raykov (1997b); however, from conceptual viewpoint it is comparable to the one described above.

An obvious advantage of ACM over coefficient α are relaxed assumptions about true scores. This means that large differences between item true variances will not lower the reliability coefficient. However, unidimensionality is still required. The other important virtue of ACM is that it allows testing the model fit. If the data fit the congeneric model well, we can be sure that estimated reliability is close to the true reliability. Additionally it is possible to test essentially τ-equivalent and parallel models.

ACM also has some drawbacks. First, the model does not implicate the most appropriate method for parameter estimation. Maximum likelihood estimation is usually used, because of its favourable inferential properties, but in general one

could also use least squares. Second, maximum likelihood estimation rests on certain assumptions (e.g., multivariate normality) which may be more or less false in practice (Bollen, 1989). The situation is especially delicate when items are dichotomous. Third, sampling theory for reliability of a composite has not been developed. So we can determine confidence intervals only for true and error variances of items and not for the reliability of the whole test. One could use bootstrap estimation here, but that would be very time-consuming.

1.3 The greatest lower bound for reliability

The last approach to be described here is the greatest lower bound (GLB) for reliability. In fact this is not really a single method but rather a theoretical concept with several possible computational approaches. The concept of GLB was introduced by Jackson and Agunwamba (1977). According to basic postulates of the classical test theory, the covariance matrix of the observed scores is a sum of the true scores covariance matrix and error scores diagonal covariance matrix:

$$\mathbf{C}_X = \mathbf{C}_T + \mathbf{C}_E \tag{7}$$

Since all of these matrices are covariance matrices, they are necessarily non-negative definite. Jackson and Agunwamba argued that the greatest lower bound to reliability can be determined by finding covariance matrix of errors with largest trace, subject to the condition that resulting matrices $\mathbf{C}_T$ and $\mathbf{C}_E$ are still non-negative definite. In other words, GLB corresponds to the lowest possible reliability coefficient which is still consistent with the data structure. Jackson and Agunwamba showed that some previously developed lower bounds, including coefficient α and split-half coefficient, can be derived subject to similar conditions, but never exceed the GLB.

In computation of GLB only diagonal elements of the target $\mathbf{C}_T$ and $\mathbf{C}_E$ have to be found because covariances of true scores are equal to covariances of observed scores and covariances of error scores are zero. In spite of this simplification the task is quite complicated. An analytical solution is generally impossible, but the best iterative solution is also not self-evident. Different approaches were suggested by Woodhouse and Jackson (1977), Bentler and Woodward (1980), and finally by ten Berge and Kiers (1991). The latter authors suggested a solution by means of the Minimum rank factor analysis (MRFA). MRFA is a factor analytical procedure which searches factor solution by minimising a certain number of smallest eigenvalues, so that both common and unique variances are non-negative. The special case when one minimises the sum of all eigenvalues, which is equivalent to maximising uniquenesses and minimising communalities, is in fact the case of

finding the error covariance matrix with largest trace which is used in determining the GLB.

Primary strength of the GLB is that it is the most accurate estimate of reliability we can find. But there are some reasons for caution, too. Procedures for finding the GLB are not very widely used, including MRFA, which is probably the most elaborated. That means that they have not been as closely examined by independent researchers as the methods discussed before. It is probably clear that complex iterative methods have much place for hidden faults (e.g. local minima, nonconvergence, accumulation of rounding error). In addition, software for performing computations is still in development and is not publicly available, which is a considerable practical problem.

Another problem is the lack of sampling theory. Generally, the model assumes that we know the population covariance matrix. The authors of MRFA have found out that there is a severe positive bias in small samples, and recommend using samples of size above 1000 (ten Berge, 1998). In many cases of psychological, social science or biomedical research it is very difficult to obtain such large samples.

From the above discussion it follows that the three described reliability estimates are not equally efficient in identifying the amount of true variance. According to their underlying assumptions, α should always be the lowest (since it is based on the most strict assumptions) and GLB the highest. It can be concluded that the main factors influencing differences in efficiency of various reliability coefficients are multidimensionality and differences in true variances. Specifically, we should expect the following.

1. The larger the differences between true variances, the less efficient will be coefficient α compared to ACM and GLB estimates. These differences should not affect the ACM estimate, since it only assumes unidimensionality.

2. The more the data depart from unidimensionality, the less efficient will be both α and ACM coefficient compared to the GLB. α and ACM estimates explicitly assume congeneric measurement and therefore the departure from this condition should affect their size more than the size of the GLB, which does not require unidimensionality.

2 Examples

2.1 Simulated data

All three methods will now be illustrated and compared using some simulated data sets.

Since this paper is not dealing with sampling issues, all covariance matrices used here are supposed to be population matrices. To compare the effectiveness of various reliability estimates, covariance matrices of true, error and observed scores were constructed. An obvious advantage of this approach is that the exact value of the reliability coefficient can be computed. Four imaginary tests, corresponding to the covariance matrices, consisted of six items each. The covariance matrices corresponded to the following measurement models:

1. Congeneric model with perfectly correlated true scores and different true and error variances. The reliability of all items was set to 0.2 and the ratio of the smallest to the largest variance was 1:3.

2. Non-congeneric model with one common factor: in this case the data were unidimensional in a sense that there was only one underlying common dimension, but the true scores shared only two thirds of common variance.

3. Non-congeneric model with two common factors: here there were two underlying dimensions, but the items were at least partly loaded on both of them to avoid unrealistically low correlations (ratio between higher and lower loading was about 1:2.5 for each variable).

4. Non-congeneric with three common factors: this model was the same as the previous one, with three common factors instead of two.

Three reliability estimates were computed for each case; ACM estimates were computed by means of LISREL program (Jöreskog and Sörbom, 1993) and GLB's were computed by MRFA2 program (Kiers, 1996). ACM estimates were computed from the covariance matrix using ML estimation. ML estimation was used since simulated variables were composed as multinormal and continuous. The estimates were then compared to the true reliability, which was about 0.60 in all cases.

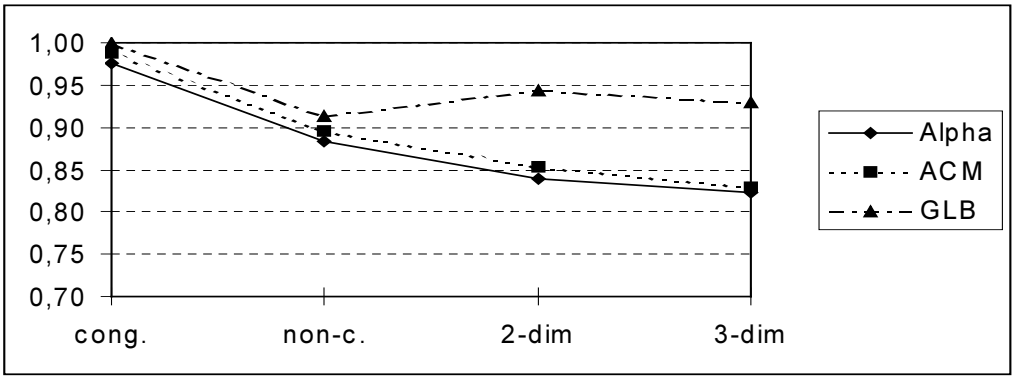

Figure 1: Reliability estimates divided by true reliability. [2]

Figure 1 shows these estimates divided by the true reliability, so that we can see what proportion of true variance do various methods identify. We can see that in the case of exactly congeneric model coefficient α was very effective in spite of unequal true score variances. The other two coefficients were almost perfectly effective.

In other three cases α and ACM estimates were very close, ACM estimate being always a bit higher. Their sizes, relative to the true reliability, diminished as the model departured more from the congeneric model. In the case of the three-dimensional structure, the size of α was only 82% of the true coefficient. We could reasonably expect the difference between α and ACM coefficients to be larger if the variability of the true score variances was greater. The size of GLB was not so much affected by multidimensionality and remained above 90% of true reliability in all cases.

2.2 Real data

Two datasets were used, both obtained on relatively large samples.

1. Personality scales: Big Five Questionnaire

The analyses were based on a representative sample of 1525 adults used for the Slovene standardisation of the Big Five Questionnaire (BFQ; Caprara, Barbaranelli, Borgogni, Bucik and Boben, 1997; see also Bucik, Boben, and Kranjc, 1997). This is a newly developed self-report instrument intended to measure five basic dimensions of personality: Energy, Agreeableness, Conscientiousness, Emotional stability and Openness. Each of the scales consists of 24 items and is divided into two subscales of 12 items each. Subscales are

[2] Note that the X-axis is not linear and that Y-axis does not begin at zero.

intended to measure somewhat different aspects of the main dimension (e.g., the subscales of Energy are Activity and Dominance). The items of BFQ are statements about typical behaviours and feelings; each tested subject rates his/her agreement with a statement on a five-point numeric scale. Since there is the same (five-point) answer scale for all items, we should not expect very large differences between item variances. However, we can expect certain degree of multidimensionality. Namely, each personality scale is composed of two subscales which measure slightly different latent traits.

Results for two scales will be shown. The first is Energy: persons with high scores are dominant, active, dynamic. The other is Openness: persons with high scores are well-informed and interested in other cultures and new experiences.

2. Attainment tests: final high-school examinations

The other dataset was obtained on the Slovene national final high-school examination in 1996[3]. First part of it are the results of the exam in mathematics. The test consisted of 14 written questions (scored on five- to seven-point scales) and an oral examination (contributing 20% to the total score). The number of candidates was 6318. The second part are results of the psychology exam. This exam consisted of an essay part (scored by maximum 45 points), 13 written questions (scored on one- to seven-point scales) and an oral examination. There were 1183 candidates taking part in this examination.

ACS estimation was again based on covariance matrix and ML estimation. ML estimation was chosen in spite of the fact that some variables were dichotomous, since many variables had too many different values to allow treatment for ordinal variables.

Table 1 shows the reliability estimates for all tests. As expected, α is the lowest and GLB is the highest in all cases. In three out of four cases, α and ACM estimate are practically the same size, the latter being just slightly higher than α. The only exception is the psychology exam, where the difference between them is 0,12. Differences between GLB and ACM estimates are mostly larger than differences between α and ACM, the largest being 0,07.

Table 1: Reliability estimates of the personality and knowledge tests.

	α	ACM	GLB
Energy	.828	.832	.889
Openness	.798	.806	.871
Mathematics	.837	.843	.864
Psychology	.652	.771	.798

[3] The data were obtained from the National examinations centre (RIC).

It was suggested previously that the differences between the reliability estimates are caused by departures from assumptions of essential τ-equivalence and congenerity. There are several ways to test hypotheses about these assumptions, but the most direct one is probably by means of confirmatory factor analysis (CFA). To test the hypothesis that measurements are congeneric, one has to fit a one-factor CFA model; the essential τ-equivalence hypotheses corresponds to one-factor CFA model with equality constraints on regression coefficients (Jöreskog, 1971). Goodness-of-fit indices can be used as indicators of validity of assumptions about respective models.

Table 2: Adjusted goodness-of-fit indices for congeneric and τ-equivalent model.

	Model	
	τ-equivalent	congeneric
Energy	.83	.85
Openness	.83	.86
Mathematics	.89	.98
Psychology	.47	.94

Table 2 presents adjusted goodness-of-fit indices (AGFI) for all measures. AGFI was chosen for the following reasons: it is not affected by the number of degrees of freedom, it is relatively easy to interpret since it is scaled to range between 0 and 1, and finally, it indicates the proportion of variance explained by the model, which is close to the classic rationale of reliability (see Jöreskog and Sörbom, 1993). Both knowledge tests, especially the mathematics test show satisfactorily good fit to the congeneric model. Both personality scales, however, do not seem to be unidimensional. This is not surprising, since each of them consists of two subscales. Besides, personality scales rarely prove to be unidimensional in practice, even if claimed to be such.

On the other hand, difference between fit indices for both models is small (0,02 and 0,03) in case of both personality scales. Neither congeneric nor essentially τ-equivalent model showed acceptably fit here. This means that the lack of fit to the essentially τ-equivalent model was induced mostly by multidimensionality and not so much by differences between true variances. Actually, one should expect true variances to be of similar size, since all items are answered on the same five-point scale. Large differences between true variances could occur only in unlikely cases if the item variances were very different (which is not the case here) or if the items had very variable reliabilities.

The psychology exam is just the opposite: its high fit index for the congeneric model suggests that the test is close to unidimensionality, but at the same time a

very low fit index for the τ-equivalent model indicates large differences among true variances. The main reason for this are obviously the extremely variable item variances due to different maximum points. Most questions required rather short answers and some were scored just zero or one, but there was also an essay included with mean score of 23 points. The largest observed item variance is 76,13 and the smallest is only 0,02 (for comparison, in case of Energy the largest and the smallest variances are 1,29 and 0,87, respectively). Equality of true variances is extremely unlikely under such conditions. A consequence of this large departure from essentially τ-equivalent model is that coefficient α severely underestimates reliability of the exam.

To illustrate how differences among the three reliability estimates are influenced by goodness-of-fit to measurement models, correlation coefficients were computed between AGFI for each model (see Table 2) and differences between pairs of reliability estimates (computed from Table 1). They are presented in Table 3.

Table 3: Correlations between AGFI and difference between reliability estimates.

	AGFI	
	τ-equivalent	congeneric
GLB - ACM	.33	-.96
ACM - α	-.99	.35

Of course, one should interpret only the relative and not the absolute size of these coefficients. We can see that better fit to congeneric model also means that ACM estimate is close to the greatest lower bound (the higher the AGFI, the lower the difference), as in case of mathematics and psychology exams. Further, if the items are not essentially τ-equivalent, the difference between α and ACM estimate will be large, as with psychology exam. However, if the fit to essentially τ-equivalent model is good (relative to the fit to congeneric model), then both α and ACM estimate will be approximately equally effective, as it is the case with both personality scales.

3 Conclusions

We have seen how relative effectiveness of the three methods for determining reliability depends on the structure of the test. The choice of the most appropriate method should ideally be guided by the test of essentially τ-equivalent and congeneric model, but unfortunately the software for performing these analyses is

not easily available to all practitioners. Some general considerations could be as follows. Choice of coefficient α is preferable, when:

- our computational facilities are limited but we need a quick estimate of reliability,
- we are interested in confidence intervals,
- we have a small sample,
- a goodness-of-fit index for essentially τ-equivalent model is satisfactory or item covariances are of similar size, which is a symptom of essential τ-equivalence.

We should use ACM estimate instead, when a goodness-of-fit index for essentially τ-equivalent model is low but the fit to congeneric model is still satisfactory. In absence of software for structural equation modelling, we can assess unidimensionality by inspecting an eigenvalue plot, which can be produced by most statistical packages. If the data are clearly multidimensional, we should try to split our scale into two or more subscales, and then compute reliability separately for each of them.

Finally, we could use GLB when:

- we have a very large sample and we wish to have maximally accurate estimate,
- we intend to perform the correction for attenuation — in this case we can use GLB even if the sample is small since the positive bias is conservative in this situation.

From the above examples it can be concluded that coefficient α is relatively accurate even if the assumption of essential τ-equivalence is moderately violated. This coefficient can thus be used safely in most practical situations. However, when the departure from essential τ-equivalence seems to be large (e.g. some variances and especially covariances are several times greater than other), one should prefer other measures of reliability.

References

[1] Bentler, P.M. and Woodward, J.A. (1980): Inequalities among lower bounds to reliability: with applications to test construction and factor analysis. *Psychometrika,* **45**, 249-267.

[2] Bollen, K.A. (1989): *Structural Equations with Latent Variables.* New York: Wiley.

[3] Bucik, V., Boben, D., and Kranjc, I. (1997): Vprašalnik BFQ in ocenjevalna lestvica BFO za merjenje "velikih pet" faktorjev osebnosti: slovenska priredba. [BFQ questionnaire and BFO rating scale for measurement of the "big five" factors of personality: Slovene adaptation.] *Psihološka obzorja,* **6**, 5-34.

[4] Caprara, G.V., Barbaranelli, C, Borgogni, L., Bucik, V, and Boben, D. (1997): *Model velikih pet: pripomočki za merjenje osebnosti. Priročnik.* [The big five model: instruments for measurement of personality. Manual.]. Ljubljana: Produktivnost.

[5] Feldt, L.S., Woodruff, D.J., and Salih, F.A. (1987): Statistical inference for coefficient alpha. *Applied Psychological Measurement,* **11**, 99-103.

[6] Jackson, P.H. and 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.

[7] Jöreskog, K.G. (1971): Statistical analysis of sets of congeneric tests. *Psychometrika,* **36**, 109-133.

[8] Jöreskog, K.G. and Sörbom, D. (1993): *LISREL 8 - User's Reference Guide.* Chicago, IL : Scientific software international.

[9] Kiers, H.A.L. (1996): MRFA2: a computer program for Minimum rank factor analysis [Computer software]. Groningen: University of Groningen.

[10] Nunnally, J.C. and Bernstein, I.H. (1994): *Psychometric Theory.* New York: McGraw-Hill.

[11] Lord, F.M. and Novick, M.R. (1974): *Statistical Theories of Mental Test Scores (2nd printing).* Reading, MA: Addison-Wesley.

[12] Raykov, T. (1997a): Scale reliability, Cronbach's coefficient alpha, and violations of essential tau-equivalence with fixed congeneric components. *Multivariate Behavioral Research,* **32**, 329-353.

[13] Raykov, T. (1997b): Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement,* **21**, 173-184.

[14] ten Berge, J.M.F. (1998): *Some recent developments in some classical psychometric problems*. Paper presented at the 9th European conference on personality, Gent, Belgium.

[15] ten Berge, J.M.F. and Kiers, H.A.L. (1991): A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika,* **56**, 309-315.

[16] Woodhouse, B. and 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.

[17] Woodward, J.A, and Bentler, P.M. (1978): A statistical lower bound to population reliability. *Psychological Bulletin,* **85**, 1323-1326.