Article

Assessing the Quality and Appropriateness of Factor Solutions and Factor Score Estimates in Exploratory Item Factor Analysis

Educational and Psychological Measurement
2018, Vol. 78(5) 762–780

DOI: 10.1177/0013164417719308
journals.sagepub.com/home/epm


Pere J. Ferrando[1] and Urbano Lorenzo-Seva[1]

Abstract

This article proposes a comprehensive approach for assessing the quality and appropriateness of exploratory factor analysis solutions intended for item calibration and individual scoring. Three groups of properties are assessed: (a) strength and replicability of the factorial solution, (b) determinacy and accuracy of the individual score estimates, and (c) closeness to unidimensionality in the case of multidimensional solutions. Within each group, indices are considered for two types of factor-analytic models: the linear model for continuous responses and the categorical-variable-methodology model that treats the item scores as ordered-categorical. All the indices proposed have been implemented in a noncommercial and widely known program for exploratory factor analysis. The usefulness of the proposal is illustrated with a real data example in the personality domain.

Keywords

exploratory item factor analysis, factor determinacy, marginal and conditional reliability, EAP-estimation, *H* index, closeness to unidimensionality

Exploratory (unrestricted) factor analysis (EFA) is a particular type of structural equation model (SEM) with latent variables. So, the degree of goodness of the model–data fit of any EFA solution can be assessed by using available procedures

[1]Universitat Rovira i Virgili, Tarragona, Spain

Corresponding Author:
Pere J. Ferrando, Research Centre for Behavioral Assessment, Universitat Rovira i Virgili, Facultad de Psicología, Carretera Valls s/n, 43007 Tarragona, Spain.
Email: perejoan.ferrando@urv.cat

intended for SEMs in general (e.g., Ferrando & Lorenzo-Seva, 2017; Yuan, Chan, Marcoulides, & Bentler, 2017). In principle, an acceptable fit is a basic requirement for judging an EFA solution as appropriate. However, the sole reliance on this requirement does not guarantee that the solution is a good one or is of practical usefulness, a point that is particularly relevant when EFA is used as a psychometric tool for item calibration and individual scoring. Indeed, it is quite possible to obtain an acceptable fit in a poorly determined solution based on low-quality items, which, in turn, yields unreliable and indeterminate factor score estimates. Also, an essentially unidimensional solution might require a multidimensional solution to be specified if the model–data fit is to be acceptable. However, this solution might well consist of additional minor and ill-defined factors of no substantive interest (e.g., Reise, Bonifay, & Haviland, 2013).

Several complementary indices have been proposed for assessing the determinacy, quality, and usefulness of psychometric FA solutions. Most of them have focused on the unidimensional case (see, e.g., Hancock & Mueller, 2000), but recently, Rodriguez, Reise, and Haviland (2016a, 2016b) have put forward a well-organized proposal in the context of bifactor FA solutions (Reise, 2012). Also, most of the indices are derived from the standard linear FA model. In this framework, most derivations are quite direct because both the item scores and the factor score estimates are linearly related to the common factors.

In practice, most item scores are discrete and bounded, so the linear FA model can only be approximately correct (at best) when they are fitted. Our position is that the linear approximation is reasonable when (a) the items have nonextreme distributions and moderate discriminating power and (b) the number of categories is relatively high (see Culpepper, 2013; Ferrando, 2009; Rhemtulla, Brosseau-Liard, & Savalei, 2012). When these conditions are not met, it is generally better to use categorical-variable-methodology factor analysis (CVM-FA). CVM-FA is briefly summarized below, but the most relevant point regarding the present developments is that the relations between the factor(s) and the observed item scores are no longer linear.

The main aim of the present article is to propose a general approach for assessing the quality, accuracy, and usefulness of a psychometric EFA application. The organization of our proposal closely follows that by Rodriguez et al. (2016a, 2016b). However, there are important differences in both scope and content. First, we focus mainly on multiple oblique solutions. Second, we consider measures based on both linear FA and CVM-FA. Third, we are not concerned with sum test scores but only with factor score estimates derived from calibration results. Finally, we propose only simple indices that will be implemented in a well-known, noncommercial EFA program, and which can be routinely used by the practitioner.

We shall now go on to summarize the starting position and scope of our proposal. We consider full psychometric applications in which FA is used for both item calibration and individual scoring. In this context, we consider that a good FA solution not only has to reach an acceptable level of goodness of model-data fit but also has

Table 1. Summary of the Indices Proposed.

Property	Linear FA	CVM-FA
Factor score determinacy and accuracy	FDI (regression based) Marginal reliability (regression based)	FDI (EAP-based) Marginal reliability (EAP-based) Individual reliabilities
Construct replicability	G-H	G-H latent G-H observed
Closeness to unidimensionality	ECV-global I-ECV IREAL-global IREAL-item	ECV-global I-ECV IREAL-global IREAL-item

Note. FA = factor analysis; CVM-FA = categorical-variable-methodology factor analysis; FDI = factor determinacy index; EAP = expected a posteriori; ECV = explained common variance; IREAL = item residual absolute loadings.

to provide (a) a clearly interpretable and strong pattern solution expected to be replicable across samples and studies and (b) factor score estimates that are determinate and accurate. The need for these strong requirements, however, should be qualified. If only the assessment of the test structure is of interest, then only requirement (a) is relevant. Requirement (b) is relevant in validity assessments based on estimated scores and, above all, in individual assessment.

In our proposal, property (a) above—the strength and replicability of a pattern or structure solution—is assessed by using extensions of Hancock and Mueller's (2000) *H* index, while property (b)—the determinacy and accuracy of the individual trait estimates—is assessed by using different determinacy and reliability indices.

Many psychometric measures were initially intended to be essentially unidimensional. However, as mentioned above, the EFAs of these measures in most cases yield multidimensional solutions in which the factor structures and derived score estimates do not reach the requirements discussed above. In this case it is quite relevant to assess how close a multidimensional solution is to a unidimensional solution, and we also propose indices to assess this issue. Overall, a summary of the present proposal is given in Table 1.

Background

Consider a test, made up of n items, that measures m traits or common factors θ_k. Let X_{ij} be the observed score of respondent i on item j. In the linear EFA model, X_{ij} is taken as a continuous-unbounded variable, and its expected score is given by

$$E(X_{ij}|\boldsymbol{\theta}_i) = \lambda_{j1}\theta_{i1} + \cdots + \lambda_{jk}\theta_{ik} + \cdots + \lambda_{jm}\theta_{im} \quad (1)$$

where the λs are the factor loadings. Both the *X*s, and the factors, θs, are scaled in a *z*-score metric (mean 0 and variance 1), so the λs are standardized loadings. For fixed $\boldsymbol{\theta}$, the *X*s become linearly independent and their conditional distributions are assumed to be normal. Furthermore, the marginal distribution of $\boldsymbol{\theta}$ is also assumed to be normal. The structural correlation matrix implied by Model (1) is

$$\mathbf{R} = \boldsymbol{\Lambda}\boldsymbol{\Phi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi} \tag{2}$$

where $\boldsymbol{\Lambda}$ is the pattern loading matrix, $\boldsymbol{\Phi}$ is the interfactor correlation matrix, and $\boldsymbol{\Psi}$ is the diagonal matrix of the item residual variances.

In the CVM-FA case, Model (1) is assumed to hold for latent response variables X^*s, normally distributed and scaled in a *z*-score metric, that underlie the observed item scores

$$E(X_{ij}^*|\boldsymbol{\theta}_i) = \lambda_{j1}\theta_{i1} + \cdots + \lambda_{jk}\theta_{ik} + \cdots + \lambda_{jm}\theta_{im} \tag{3}$$

Furthermore, the observed scores are assumed to arise as a result of a step function governed by $c - 1$ thresholds: $\tau_1, \ldots, \tau_{c-1}$ where c is the number of response categories

$$\begin{aligned} &X = i \Leftrightarrow \tau_{i-1} < X^* < \tau_i \\ &-\infty = \tau_0 < \tau_1 \cdots < \tau_{c-1} < \tau_c = +\infty \end{aligned} \tag{4}$$

Under the conditions described so far, the CVM-EFA implied correlation structure is that of Equation (2) in which $\mathbf{R}$ is now the interitem polychoric correlation matrix. With reparameterization, the CVM-EFA model becomes the item response theory (IRT) multidimensional two-parameter normal-ogive model for the binary case and the normal-ogive multidimensional graded response model for more than two ordered categories (see, e.g., Ferrando & Lorenzo-Seva, 2013, or McDonald, 1999). Here we shall mainly use the FA parameterization. However, some IRT results will also be used when the CVM-based indices are derived.

In the conventional EFA scenario considered here, the linear and the CVM models are fitted by using a random-regressors two-stage estimation approach (McDonald, 1982). In the first stage (calibration), the structural item parameters in (2) and (4) are estimated. In the second stage (scoring), the item parameter estimates are taken as fixed and known, and used to estimate the individual trait levels for each respondent. We shall not consider here specific calibration procedures. However, in the scoring stage we shall consider only Bayes Expected a Posteriori (EAP) score estimates. The main reason for this choice is that these scores have the highest correlations with the common factors they measure (e.g., Mulaik, 2010). This is a basic property in some of the indices proposed here and considerably simplifies many of the present developments.

In the linear EFA model (1), and under the conditional and prior normality assumptions discussed above, the EAP point factor score estimates are known in FA terminology as ‘‘regression factor scores’’ and were originally proposed by

Thurstone (1935). The term ''factor scores,'' however, might lead to confusion between the latent factor scores (which are unknown) and the score estimates. For this reason we shall continue using the terminology ''factor score estimates.'' Strictly speaking, however, it should be noted that the term ''estimates'' is not correct in the usual statistical sense because there are no ''true'' parameter values to be approximated by these estimates (see Maraun, 1996).

In the general oblique case, regression factor score estimates can be obtained in closed form as (Thurstone, 1935)

$$EAP(\boldsymbol{\theta}_i) = \boldsymbol{\Phi}\boldsymbol{\Lambda}'\mathbf{R}^{-1}\mathbf{X}_i = \mathbf{S}'\mathbf{R}^{-1}\mathbf{X}_i \tag{5}$$

where $\mathbf{X_i}$, of dimension $n \times 1$ is the vector containing the standardized item scores of respondent i, and $\mathbf{S}$, of dimension $n \times m$ is the structure matrix whose elements are the item–factors correlations.

In the case of CVM-EFA, the EAP point estimate of $\boldsymbol{\theta}_i$ for the k dimension (θ_{ik}) cannot be obtained in closed form, and is obtained via the general definition:

$$EAP(\theta_{ik}) = \frac{\int_{\boldsymbol{\theta}} \theta_k L(\mathbf{x_i}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} L(\mathbf{x_i}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{6}$$

where $g(\boldsymbol{\theta})$ is the joint multivariate prior density of $\boldsymbol{\theta}$ and L is the likelihood of $\mathbf{x}_i$ which can be written generically as

$$L(\mathbf{x_i}|\boldsymbol{\theta}_i) = \prod_{j=1}^{n} P(X_{ij}|\boldsymbol{\theta}_i) \tag{7}$$

And the generic expression $P(X_j|\boldsymbol{\theta})$ denotes the conditional probability assigned to a specific item score for fixed $\boldsymbol{\theta}$.

The diagonal elements of the posterior (error) covariance matrix are given by

$$PSD^2(\theta_{ik}) = \frac{\int_{\boldsymbol{\theta}} (\theta_k - EAP(\theta_{ik}))^2 L(\mathbf{x_i}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\boldsymbol{\theta}} L(\mathbf{x_i}|\boldsymbol{\theta}) g(\boldsymbol{\theta}) d\boldsymbol{\theta}} \tag{8}$$

where PSD means posterior standard deviation.

As mentioned above, the EAP estimator has minimum mean squared error, so it cannot be improved upon in terms of average accuracy (e.g., Bock & Mislevy, 1982). At any θ_{ik} level (except the population mean), however, it is inwardly biased (i.e., regressed toward the mean), and this occurs in both the linear case (Krijnen, Wansbeek, & ten Berge, 1996) and the CVM case (Bock & Mislevy, 1982). As the number of items increases, the likelihood gradually dominates the prior, the likelihood and the posterior density become virtually indistinguishable, and the EAP estimate approaches conditional unbiasedness (Bock & Mislevy, 1982). We shall refer to this result with the statement that *asymptotically* the EAP estimates are conditionally unbiased.

Determinacy and Reliability of the Factor Score Estimates

In EFA, the factor indeterminacy problem is the result that more than one set of factor score estimates can be constructed that are consistent with a given correlational structure with the form (2). This problem arises because the number of common and unique factors exceeds the number of items and has generated a considerable amount of controversy in the FA literature (see, e.g., Grice, 2001; Maraun, 1996; Mulaik, 2010).

From a practical perspective, the most usual way to address the problem stated above is to assess the degree of indeterminacy of the score estimates (e.g., Grice, 2001). According to Cliff (1977), consistency of person ordering is the primary goal of individual assessment, and a high degree of indeterminacy implies precisely that respondents cannot be consistently ordered along the trait continuum. It also implies that the validity relations between the factor score estimates and relevant criteria are also indeterminate. Given the practical relevance of the problem, the degree of indeterminacy should be routinely assessed in FA studies of the type considered here, but unfortunately, this does not appear to be the case (Grice, 2001).

Of the various indices that quantify the extent to which the scores are indeterminate (Guttman, 1955), the most common is possibly the correlation between the factor score estimates and the levels on the latent factors they estimate (Beauducel, 2011). We shall denote this index by $\rho_{(\hat{\theta}\theta)}$ and name it ''factor determinacy index'' (FDI). When the FDI value is near one, the factor score estimates are good proxies for representing the latent factor scores, and the different factor score estimates that are compatible with the given structure are also highly correlated with one another (Guttman, 1955). As for reference values, Gorsuch, (1983) considered that FDI values around 0.80 will be adequate for research purposes. However, if the scores are to be used for individual assessment, a value of 0.90 may be a minimal requirement (Grice, 2001; Rodriguez et al., 2016a).

We shall first consider linear FA. The FDI estimates based on the regression scores are the diagonal elements of the $m \times m$ matrix:

$$\left[\boldsymbol{\Phi}\boldsymbol{\Lambda}'\mathbf{R}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi})\right]^{1/2} = \left[\mathbf{S}'\mathbf{R}^{-1}\mathbf{S})\right]^{1/2} \tag{9}$$

(e.g., Beauducel, 2011). As mentioned above, the FDIs in (9) are the highest possible of all the types of factor score estimates.

The unidimensional case is useful for understanding the determinants of the FDI values. In this case, the FDI is obtained as

$$\rho_{(\hat{\theta}\theta)} = (\boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda})^{1/2} = \frac{1}{\sqrt{1 + \frac{1}{\sum_{j=1}^{n} \frac{\lambda_j^2}{\sigma_{\varepsilon j}^2}}}} \tag{10}$$

The term $\sum_{j=1}^{n} \lambda_j^2 \Big/ \sigma_{\varepsilon j}^2$ in (10) is the (constant) amount of information in the linear FA model (Ferrando, 2009; Mellenbergh, 1996). Clearly, the degree of determinacy

depends on (a) the number of items and the signal-to-noise ratios between the squared loadings and the residual variances. In the standardized modeling considered here, the residual variances depend only on the loadings, so test length and the magnitude of the loadings are the sole determinants of FDI.

The square of $\rho_{(\hat{\theta}\theta)}$ is one of the standard definitions of a reliability coefficient (Brown & Croudace, 2015; Mellenbergh, 1996). So, by this definition, the squared values of the FDI estimates obtained in (9) are interpreted as the reliabilities of the corresponding factor score estimates.

We turn now to CVM-FA. Reliability estimates based on the $\rho^2_{(\hat{\theta}\theta)}$ definition (and, therefore, on the corresponding FDI estimates) have received some attention in the IRT literature (Green, Bock, Humphreys, Linn, & Reckase, 1984; Samejima, 1977). To derive the FDIs in this case, we shall write the EAP estimated score for individual i in factor k as

$$\hat{\theta}_{ik} = \theta_{ik} + \delta_{ik} \tag{11}$$

(see, e.g., Samejima, 1977). If the estimator in (11) is conditionally unbiased, then by standard covariance algebra, it follows that the FDI could be obtained as

$$\rho_{(\hat{\theta}_k\theta_k)} = \sqrt{\frac{Var(\hat{\theta}_k) - Var(\delta_k)}{Var(\hat{\theta}_k)}} = \sqrt{\frac{Var(\hat{\theta}_k) - E(PSD^2(\theta_{ik}))}{Var(\hat{\theta}_k)}} \tag{12}$$

And its squared value is the corresponding reliability estimate. This estimate is an empirical estimate (Brown & Croudace, 2015) which uses (a) the variance of the EAP scores and (b) the average of the squared PSDs both obtained in the calibration sample. However, unlike the linear estimate (9), which is correct for any number of items, (12) is only asymptotically correct, because, as discussed above, the estimator (11) is only asymptotically unbiased. As discussed below, in very short tests, we expect (12) to be somewhat upwardly biased.

A conditional or individual reliability estimate (Green et al., 1984, Raju, Price, Oshima, & Nering, 2007) can further be obtained as

$$\hat{\rho}(\theta_{ik}) = \frac{Var(\hat{\theta}_k) - PSD^2(\theta_{ik})}{Var(\hat{\theta}_k)} \tag{13}$$

So the reliability marginal estimate (i.e., the squared value in 12) is the average of the individual estimates in (13). We propose to obtain the distribution of these individual estimates as auxiliary information that complements the information provided by the marginal estimate. To see the interest of this additional measure, consider that an acceptable marginal reliability estimate is still compatible with the presence of a non-negligible proportion of respondents that cannot be accurately measured.

To close this section, we note that Beauducel and Hilger (2017) considered a scoring schema that is half way between (5) and (6), and derived unbiased FDIs (and marginal reliability estimates) based on the resulting score estimates. Specifically they

considered obtaining linear regression estimates of the form (5) in which **S** and **R** were based on the CVM but $\mathbf{X_i}$ contained the observed categorical scores. We shall not consider this approach in the present proposal, but it would be of interest in the future to assess how the resulting FDI and reliability estimates behave in comparison to those proposed here.

Construct Replicability

Hancock and Mueller (2000) and Hancock (2001) proposed an index to assess the extent to which a factor is well represented by a set of items. This general concept comprises several properties (mainly, the quality of the items as indicators of the factor, and the replicability of the factor solution across studies). Hancock and Mueller (2000) labeled their index *H*, and used the term ''construct reliability.''Rodriguez et al. (2016b) renamed it as ''construct replicability,'' which is the name we shall use here. The initial proposal considered only the unidimensional case, and, using the present notation, can be written as

$$H = (\boldsymbol{\lambda}'\mathbf{R}^{-1}\boldsymbol{\lambda}) = \frac{1}{1 + \frac{1}{\sum_{j=1}^{n} \frac{\lambda_j^2}{\sigma_{\varepsilon j}^2}}} \tag{14}$$

Essentially (14) measures the maximal proportion of the variance of the factor that can be accounted for by its indicators. So, *H* is the squared correlation between the factor and an optimal composite of its indicator scores, or in other words, the squared multiple correlation between the factor and its indicators. We note that (14) is the square of the FDI measure in (10) and, therefore, the reliability estimate we propose here for the unidimensional linear model. This result is only to be expected given that the regression factor score estimates are the optimal linear composite that maximizes the multiple correlation.

In the general oblique case, the multiple correlations between the factors and their indicators are obtained as the squared diagonal elements of the matrix (9) (e.g., Mulaik, 2010, Equation 13.16)

$$G - H = diag\left[\boldsymbol{\Phi}\boldsymbol{\Lambda}'\mathbf{R}^{-1}\boldsymbol{\Lambda}\boldsymbol{\Phi})\right] = diag\left[\mathbf{S}'\mathbf{R}^{-1}\mathbf{S}\right] \tag{15}$$

We propose to use these elements as generalized *H* indices (denoted by *G-H*) for multidimensional oblique solutions. To justify this choice, we note that, in terms of structural coefficients, *G-H* has the same basic properties as has the original *H* in terms of standardized loadings. First, it is not affected by the sign of the structural coefficients. Second, its value is always at least as large as the largest squared structural coefficient (Yule, 1907, Equation 17). Finally, the addition of an indicator will always increase the existing *G-H* value or leave it same. The maximum value of *G-H* is 1 and will occur when one of the indicators has a perfect correlation with the common factor. Initially, Hancock and Muller (2000) proposed 0.70 as a minimal

reference value because the factor was well represented. Rodriguez et al. (2016b) raised it to 0.80. For the *G-H* indices proposed here, the 0.80 cutoff also seems to be reasonable.

In summary, if the *G-H* conceptualization is accepted, it follows that, in the linear model and when regression factor score estimates are considered, the measures of determinacy, reliability, and construct replicability are all obtained from the same basic expression. So, the squared FDIs can be interpreted as both the reliabilities of the regression factor score estimates and the squared multiple correlations between the item scores and the common factors (i.e., generalized *H* measures).

We turn now to the CVM-FA where the relations are more complex. Consider first that (15) is computed by using (a) the calibration estimates obtained from fitting a CVM-FA solution and (b) the interitem polychoric correlation matrix. The diagonal elements of (15) now become the multiple correlations between the factors and the continuous latent response variables that underlie the observed item scores. We shall label the index proposed so far as *G-H-latent*.

The multiple correlations between the factors and the observed item scores are necessarily lower than the corresponding *G-H-latent* values due to (a) the nonlinearity of the item-factor regressions and (b) attenuation for coarse grouping. They can be predicted from the CVM-FA solution as follows. First, **R** can be directly estimated via the product moment interitem correlation matrix. Second, the elements of **S** in *G-H-latent* are the (polyserial) item–factor correlations. So, the product–moment item–factor correlations can be predicted from the elements of **S** by using the relation between the polyserial and the product-moment correlation (e.g., Olsson, Drasgow, & Dorans, 1982)

$$\rho(X_j, \theta_k) = \frac{\rho(X_j^*, \theta_k) \sum_{u=1}^{c-1} \phi(\tau_u)}{\sqrt{Var(X_j)}} \tag{16}$$

where ϕ is the ordinate of the standard normal distribution. The resulting measure is denoted by *G-H-observed* and, when compared to *G-H-latent*, quantifies the predicted loss of information and construct replicability that will occur if the item scores are treated as continuous-unbounded variables and fitted with the linear EFA model. We believe that this information is relevant to deciding which model is the most reasonable for a given analysis: if the differences between *G-H-latent* and *G-H-observed* are minor, the simpler linear model could be considered.

Finally, we should point out that the equivalence between the reliabilities of the factor score estimates and the generalized *H* measures does not hold in the CVM case. *G-H-latent* can be viewed as the hypothetical reliability that the regression scores would have in model (3) if the underlying latent response variables were available. Indeed, this is not the case, and the EAP estimates in (6) are obtained from the pattern of observed scores, as specified in Equation (7).

Closeness to Unidimensionality

A review of many reported oblique solutions suggests that they are compatible with an essentially unidimensional solution (Reise, Cook, & Moore, 2015; Reise et al., 2013). Furthermore, according to the proposal made here, for an oblique solution to be justifiable and useful, all the proposed factors have to be well defined and replicable (in terms of *G-H*) and lead to determinate and reliable factor score estimates. We suspect that this is not the case in most applications. So, given these results, it seems necessary to assess the extent to which an oblique EFA solution is close to unidimensionality, and interpretable in these terms. In this assessment, it should also be considered that forcing a unidimensional solution on data that is clearly multidimensional can lead to biased results in which the single fitted factor does not reflect a unitary construct but is, essentially, a weighted composite of the different factors.

A simple and informative index that assesses closeness to unidimensionality has been proposed in slightly different variants for the linear FA model (see, e.g., Rodriguez et al., 2016a, 2016b). Here we propose using the version by ten Berge and Kiers (1991) based on minimum rank factor analysis (MRFA). For a unidimensional solution, MRFA produces a reduced correlation matrix (with communalities in the main diagonal) so that the sum of its eigenvalues except the first one is the smallest possible. Conceptually this is equivalent to obtaining a canonical factor solution (e.g., Harman, 1962) in $n-1$ factors in which the sum of the squared loadings on the first factor is the maximum possible and the sum of the squared loadings on the remaining $n-2$ factors is the smallest possible. A natural index in this setting is the explained common variance (ECV) index, which in terms of factor loadings is given by

$$ECV = \frac{\sum_{j} \lambda_{j1}^2}{\sum_{j} \lambda_{j1}^2 + \sum_{j} \lambda_{j2}^2 + \cdots + \sum_{j} \lambda_{jn-1}^2} \tag{17}$$

Stucky, Thissen, and Edelen (2013) proposed that ECV should also be computed at the single item level j, and that the resulting index be labelled I-ECV. Here we propose that this index (as derived from 17 in our case) also be used as an auxiliary measure useful for detecting the items that most contribute to the departure from unidimensionality.

Essentially, (17) measures the relative magnitude of the squared loadings on the first MRFA factor with respect to the magnitude of the full set of squared loadings on the complete MRFA solution in $n-1$ factors. So, in principle, the index can be directly computed from the linear and CVM solutions (although the interpretation in terms of explained common variance is different). We also note that the index can be computed with no need to specify a particular alternative solution in terms of structure or number of factors. Finally, regarding cutoff values, it has been proposed that ECV cutoff values should be in the range 0.70 to 0.85 if it is to be concluded that a

solution is essentially unidimensional (Green et al., 1984; Rodriguez et al., 2016a, 2016b; Stucky et al., 2013).

As defined above, ECV essentially measures the dominance of the first MRFA factor over the other factors. However, a clear dominance is still compatible with potentially biasing multidimensionality (e.g., Reise et al., 2015). To address this issue we propose that an auxiliary, model-independent, index also be used. Consider the pattern with the first and second factors of the MRFA solution described above. This pattern represents the most general common factor that can be obtained from the data plus an orthogonal residual second factor. We propose to use the absolute loadings on the second MRFA factor as measures of departure from unidimensionality at the item level, and denote them as ''item residual absolute loadings'' (IREAL). The average of these loadings can then also be used as a general measure of departure from unidimensionality. Note that these indices address the basic concept of unidimensionality that the residual loadings must be negligible regardless of the magnitude of the loadings on the dominant factor (e.g., Green et al., 1984). So, if their values are consistently low, no substantial bias can possibly be expected if a unidimensional solution is fitted. With regard to threshold values, the most common rule of thumb for judging a loading as salient is 0.30 (e.g., Grice, 2001), and tentatively, we propose this criterion as a rough initial reference.

Implementation

All the indices proposed in this article have been implemented in version 10.5 of the program FACTOR (Ferrando & Lorenzo-Seva, 2017), a well-known, free exploratory factor analysis program that can be downloaded at http://psico.fcep.urv.cat/utilitats/factor/. Indices of determinacy, reliability, and construct replicability are provided as default output for both linear and CVM solutions. Indices of closeness to unidimensionality are provided as default when a unidimensional solution is requested, and are optional otherwise. All the proposed indices are relatively simple to implement. However, as Grice (2001) noted in the context of factor score assessment, no commercial or widely available programs appear to provide this type of index.

Hancock and Mueller (2000) considered that it was important to report confidence intervals (CIs) for *H*, and proposed that they be derived with Bootstrap resampling. We believe that this point is also relevant for all the indices proposed here. In principle, CIs for some of the indices based on linear FA, mainly FDIs, marginal reliabilities, and G-H indices, could be analytically approximated by using the delta method (see, e.g., Raykov, 2002). For the remaining indices, however, an analytical treatment appears to be very complex. For this reason, we decided to implement a unified treatment in FACTOR in which bootstrap-based confidence intervals are available for all the indices proposed here. The 90%, 95%, and 99% confidence intervals available are (a) percentile intervals and (b) bias-corrected percentile intervals. The number of bootstrap samples can be defined by the user in the range [500, 3,000].

Illustrative Example

The real-data study in this section is based on a Spanish version of Buss and Perry's (1992) aggression questionnaire (AQ; Vigil-Colet, Lorenzo-Seva, & Morales-Vives, 2015). The AQ is a multidimensional questionnaire made up of 5-point Likert-type items intended to measure different related dimensions of aggression. For the present illustration, we chose a subset of 20 items that were expected to measure two factors: physical aggression (PA; 7 items) and nonphysical aggression (NPA; 13 items). The indicators, however, were not expected to be so factorially pure that an independent-cluster solution could be specified. So, an unrestricted solution was fitted instead. The questionnaire was administered to a sample of 538 secondary school students aged between 12 and 17 years. Data were kindly supplied by Dr. A. Vigil-Colet.

Descriptive analysis of the item scores showed that the distributions were generally not extreme and that linear EFA could be considered a reasonable approach. To illustrate all the procedures proposed here, both linear EFA and CVM-EFA solutions were fitted to the data. In both cases a two-factor solution was fitted by using robust unweighted least squares estimation as implemented in FACTOR.

Goodness of model–data fit was assessed by using both the conventional approach and the recent proposal by Yuan et al. (2017) based on equivalence testing. So far, the latter approach has only been fully developed for the root mean square error of approximation (RMSEA) and comparative fit index (CFI) measures based on the linear model, so we have only used it in this case.

For both models, goodness of fit results are in the upper panel of Table 2. The RMSEA and CFI measures are based on the second-order (mean and variance) corrected chi-square statistic proposed by Asparouhov and Muthen (2010). Overall, the fit based on the conventional approach can be considered to be acceptable and quite similar in both solutions. Equivalence-testing results for linear FA also suggests that the fit of the model is acceptable.

The canonical pattern was then rotated using the Promin criterion (Lorenzo-Seva, 1999), and the solutions are in table 2 with the dominant loadings boldfaced. The estimated inter-factor correlations were $\phi = 0.50$ (linear) and $\phi = 0.53$ (CVM).

As Table 2 shows, none of the solutions have an independent-cluster structure. However, they are quite clear: Bentler's simplicity indices are 0.997 in both linear and CVM, and the overall congruence between the linear and the ECV solution is 0.999. Overall, (a) the factors can be well distinguished, (b) the solution agrees with the ''a priori'' hypothesis, and (c) the linear and CVM patterns are very similar.

The calibration estimates were taken as fixed and known, and EAP factor score estimates, and PSDs, were obtained. In the CVM case, the prior for $\boldsymbol{\theta}$ was specified as bivariate standard normal with correlations of 0.50 (linear) and 0.53 (CVM) (see Ferrando & Lorenzo-Seva, 2016).

The results about the determinacy and accuracy of the EAP scores are in the upper rows of Table 3. For linear FA, the determinacies are acceptable for both factors, suggesting that the factor score estimates reflect quite univocally the latent levels they attempt to estimate. And the estimated reliabilities are appropriate for most

Table 2. Bidimensional EFA Results for the Illustrative Example.

(a) Goodness-of-Fit Results

	RMSEA	95% CI RMSEA	*T*-size RMSEA	CFI	*T*-size CFI	GFI	*Z*-RMSR
Linear	.054	(.051; .055)	.062 (fair)	.97	.953 (close)	.98	.052
CVM	.056	(.051; .057)		.97		.98	.058

(b) Promin Rotated Pattern

	Linear FA		CVM-FA	
Item	θ_1	θ_2	θ_1	θ_2
1	.035	**.618**	.040	**.715**
2	**.273**	.075	**.266**	.066
3	**.334**	.097	**.342**	.108
4	**.485**	.041	**.499**	.055
5	−.093	**.842**	−.094	**.892**
6	.341	.089	**.354**	.083
7	**.500**	.008	**.535**	−.001
8	−.055	**.692**	−.090	**.753**
9	**.669**	.012	**.719**	−.006
10	**.519**	.057	**.552**	.051
11	**.719**	−.211	**.757**	−.223
12	−.025	**.723**	−.032	**.771**
13	**.426**	.144	**.470**	.170
14	**.529**	−.036	**.576**	−.044
15	−.031	**.792**	−.025	**.865**
16	**.591**	.099	**.616**	.129
17	**.680**	−.140	**.730**	−.154
18	.119	**.533**	.132	**.605**
19	**.395**	.054	**.406**	.075
20	.270	**.357**	.292	**.403**

Note. RMSEA = root mean square error of approximation; CI = confidence interval; CFI = comparative fit index; GFI = goodness of fit index; RMSR = root mean square residual; FA = factor analysis; CVM-FA = categorical-variable-methodology factor analysis. Values in boldface indicate the dominant loading.

applications, although perhaps a little low for accurate individual assessment. As for the CVM-FA, the determinacy and reliability results are virtually the same as the ones obtained from the linear model for the second factor, but are clearly higher for the longer NPA factor.

Table 4 shows the within-and-between-model (linear vs. CVM) correlations between the factor score estimates. Furthermore, the correlations corresponding to the same factor measured with the different models were corrected for unreliability by using the marginal reliability estimates in Table 3, which are again displayed in the main diagonal of Table 4. Results can be summarized as follows. First, the disattenuated correlations are 1 for both factors, which again suggests that the

Table 3. Score Accuracy and Construct Replicability Results.

Index	Linear FA		CVM-FA	
	F1(NPA)	F2(PA)	F1(NPA)	F2(PA)
FDI	.921 (.908; .935)	.936 (.924; .946)	.954 (.939; .966)	.936 (.897; .965)
Marginal reliability	.849 (.824; .864)	.876 (.853; .894)	.912 (.882; .934)	.876 (.805; .931)
			Latent	Latent
G-H	.849 (.824; .864)	.876 (.853; .894)	.876 (.848; .889)	0.919 (.899; .930)
			Observed	Observed
			.865 (.841; 0.881)	.832 (.812; .849)

Note. FA = factor analysis; CVM-FA = categorical-variable-methodology factor analysis.

Table 4. Correlations Among the Factor Score Estimates With the Marginal Reliabilities in the Main Diagonal.

	$\hat{\theta}_{1L}$	$\hat{\theta}_{2L}$	$\hat{\theta}_{1CVM}$	$\hat{\theta}_{2CVM}$
$\hat{\theta}_{1L}$	.849			
$\hat{\theta}_{2L}$	.558	.876		
$\hat{\theta}_{1CVM}$	.951	.503	.912	
$\hat{\theta}_{2CVM}$	.574	.945	.556	.876

linear-based and the CVM-based score estimates measure the same factors. Second, the interfactor correlation estimates, both within and between models, agree quite well with the structural interfactor correlation estimates reported above. This second result provides more support for the FDI results above that the factor score estimates are good proxies for the corresponding latent factor scores.

Figure 1 shows the distribution of the individual reliabilities for both factors in the CVM-FA. It seems clear that Factor 1 not only has a higher marginal reliability but is also able to accurately measure most of the respondents. In contrast, although the estimated marginal reliability of Factor 2 is only a bit lower, it will provide poor measurement precision for many respondents.

Construct replicability indices and the confidence intervals are in the lower rows of Table 3. In all cases they are acceptable, which suggests that in both linear and CVM solutions both factors are well defined and so the solution is expected to remain stable across studies. In the linear case, the *G-H* values are the same as the reliability estimates, as discussed above, and they reasonably agree with the G-H-Observed values predicted from the CVM-FA. As expected, the GH-latent values are the highest for both factors, reflecting the result that the factors are better defined by the

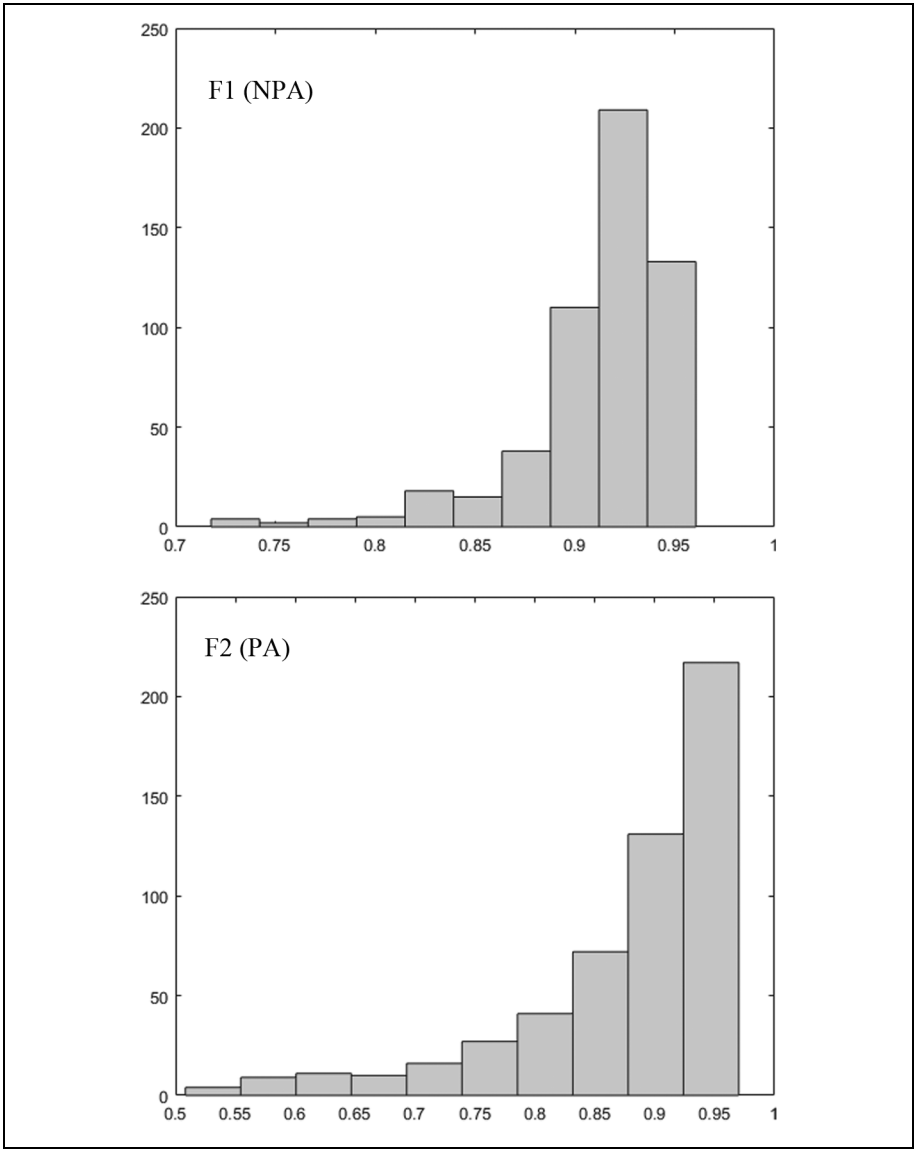

Figure 1. Distribution of the individual reliabilities estimates of the F1 and F2 scores, CVM-FA.

underlying responses than by the observed item scores. Finally, we note that the CVM-based marginal reliability for the PA factor is below the corresponding G-H value, which seems reasonable. However, this is not the case for the NPA factor, which suggests that the marginal reliability estimate for this factor is possibly a little too optimistic.

Finally, we summarize the closeness-to-unidimensionality results. The ECV values and 95% confidence intervals were the following: 0.738 (0.703; 0.769) for the linear model and 0.754 (0.721; 0.789) for the CVM model. And, for 5 items (the same in both models), the I-ECV values were below 0.70. As for the IREAL values, the averages were .276 (linear FA) and .291 (CVM FA), and 8 items (linear FA) and 9 items (CVM FA) had values above .30. Overall, and in both models, it would be marginally acceptable to consider that the AQ items measure a general common dimension of aggression. However, given that the bidimensional solution is clear, replicable, and leads to accurate and reliable factor score estimates for both factors, the oblique solution seems to be the most appropriate in this case.

Discussion

The main purpose of this article was to propose and implement a series of auxiliary indices designed to judge the quality and usefulness of FA solutions intended for psychometric applications. Our idea was to propose simple indices that could be provided as the standard output of an FA program requiring minimal specifications by the user. Overall, we believe that this purpose has been achieved, and that the proposal is potentially useful for practitioners. However, some issues deserve further discussion.

The first of these issues is the relevance and scope of the contribution. For decades, the dominant view regarding item FA has been that confirmatory FA is the way to go, while EFA is at best a rough precursor that can be useful only in the preliminary stages of the analysis (see, e.g., Ferrando & Lorenzo-Seva, 2017). In principle, we do not agree with this view and, like Cattell (1986), believe that most items are inherently complex and that unrestricted FA is the most natural and flexible approach for calibrating and scoring them. This is not an isolated opinion. In recent times there has been growing discontent among practitioners regarding the unnecessarily strong restrictions of strict confirmatory solutions more flexible methods have been on the rise (e.g., Marsh, Morin, Parker, & Kaur, 2014).

With regard to the scope, we believe it is considerable. In the illustrative example, we have purposely considered the less restricted form of EFA based on analytical rotations. However, the procedures proposed here can also be used with more restricted approaches based on Procrustres transformations against fully specified or semispecified targets, which are also available in FACTOR (e.g., Ferrando & Lorenzo-Seva, 2013).

On the practical level, we have implemented CIs based on bootstrap resampling for all the proposed indices. They seem to work well but are rather time-consuming. So, perhaps an approach in which approximate CIs are obtained analytically for the indices in which this approach is feasible, and using Bootstrap for the remaining ones would be the best option. This point is left for future research.

On the methodological level, the proposal made here is mostly based on results that are known in the psychometric or statistical literature. However, the novelty is that this is the first time so many of these results have been used in the present

context. We are not aware of generalized *H* indices being used in oblique solutions, or that they are interpreted differently in linear and the CVM models.

To summarize, we acknowledge that the proposal has its share of limitations and points that deserve further study. While factor indeterminacy and reliability estimates are correct for any number of items in the linear case, the empirical estimates in the CVM case are only asymptotically correct, and probably biased in short tests. Furthermore, this bias might well depend on the estimation method that is chosen for calibrating the items (see Beauducel & Hilger, 2017). So, the potential improvement of these estimates is an issue that warrants further research. More generally, if the procedures proposed here are to be used correctly, sensible and well-established reference values need to be provided for all the indices. This point is particularly relevant for the IREAL index in which only a rough rule of thumb has been tentatively proposed as a cutoff. Overall, further intensive research based on both simulation and real data, as well as further statistical developments are needed if reference values are to be improved.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported by a grant from the Catalan Ministry of Universities, Research and the Information Society (2014 SGR 73) and by a grant from the Spanish Ministry of Economy and Competitivity (PSI2014-52884-P).

References

Asparouhov, T., & Muthen, B. (2010, May 3). *Simple second order chi-square correction* (Unpublished manuscript). Retrieved from https://www.statmodel.com/download/WLSMV_new_chi21.pdf

Beauducel, A. (2011). Indeterminacy of factor scores in slightly misspecified confirmatory factor models. *Journal of Modern Applied Statistical Methods*, *10*, 583-598.

Beauducel, A., & Hilger, N. (2017). The determinacy of the regression factor score predictor based on continuous parameter estimates from categorical variables. *Communications in Statistics—Theory and Methods*, *46*, 3417-3425.

Beauducel, A., & Hilger, N. (2017). On the bias of factor score determinacy coefficients based on different estimation methods of the exploratory factor model. *Communications in Statistics—Simulation and Computation*, *46*, 6144-6154.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431-444.

Brown, A., & Croudace, T. (2015). Scoring and estimating score precision using multidimensional IRT. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response*

theory modeling: Applications to typical performance assessment (pp. 307-333). New York, NY: Routledge.

Buss, A. H., & Perry, M. (1992). The aggression questionnaire. *Journal of Personality and Social Psychology*, *63*, 452-459.

Cattell, R. B. (1986). The psychometric properties of tests: Consistency, validity and efficiency. In R. B. Cattell & R. C. Johnson (Eds.), *Functional psychological testing* (pp. 54-78). New York, NY: Brunner/Mazel.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika*, *42*, 375-399.

Culpepper, S. A. (2013). The reliability and precision of total scores and IRT estimates as a function of polytomous IRT parameters and latent trait distribution. *Applied Psychological Measurement*, *37*, 201-225.

Ferrando, P. J. (2009). Difficulty, discrimination, and information indices in the linear factor analysis model for continuous item responses. *Applied Psychological Measurement*, *33*, 9-24.

Ferrando, P. J., & Lorenzo-Seva, U. (2013). *Unrestricted item factor analysis and some relations with item response theory* (Technical report). Department of Psychology, Universitat Rovira i Virgili, Tarragona. Retrieved from http://psico.fcep.urv.es/utilitats/factor

Ferrando, P. J., & Lorenzo-Seva, U. (2016). A note on improving EAP trait estimation in oblique factor-analytic and item response theory models. *Psicologica*, *37*, 235-247.

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*, 236-240.

Gorsuch, R. L. (1983). *Factor analysis*. Hillsdale, NJ: LEA.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptative tests. *Journal of Educational Measurement*, *21*, 347-360.

Grice, J. W. (2001). Computing and evaluating factor scores. *Psychological Methods*, *6*, 430-450.

Guttman, L. (1955). The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, *8*, 65-81.

Hancock, G. R. (2001). Effect size, power, and sample size determination for structured means modeling and MIMIC approaches to between-groups hypothesis testing of means on a single latent construct. *Psychometrika*, *66*, 373-388.

Hancock, G. R., & Mueller, R. O. (2000). Rethinking construct reliability within latent variable systems. In R. Cudek, S. H. C. duToit & D. F. Sorbom (Eds.), *Structural equation modeling: Present and future* (pp. 195-216). Lincolnwood, IL: Scientific Software.

Harman, H. H. (1962). *Modern factor analysis* (2nd ed.). Chicago, IL: University of Chicago Press.

Krijnen, W. P., Wansbeek, T., & ten Berge, J. M. F. (1996). Best linear predictors for factor scores. *Communications in Statistics—Theory and Methods*, *25*, 3013-3015.

Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, *34*, 347-356.

Maraun, M. D. (1996). Metaphor taken as math: Indeterminacy in the factor analysis model. *Multivariate Behavioral Research*, *31*, 517-538.

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology*, *10*, 85-110.

McDonald, R. P. (1982). Linear versus models in item response theory. *Applied Psychological Measurement*, *6*, 379-396.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: LEA.

Mellenbergh, G. J. (1996). Measurement precision in test score and item response models. *Psychological Methods*, *1*, 293-299.

Mulaik, S. A. (2010). *Foundations of factor analysis*. Boca Raton, FL: CRC Press.

Olsson, U., Drasgow, F., & Dorans, N. J. (1982). The polyserial correlation coefficient. *Psychometrika*, *47*, 337-347.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*, 169-180.

Raykov, T. (2002). Analytic estimation of standard error and confidence interval for scale reliability. *Multivariate Behavioral Research*, *37*, 89-103.

Reise, S. P. (2012). The rediscovery of bifactor measurement models. *Multivariate Behavioral Research*, *47*, 667-696.

Reise, S. P., Bonifay, W. E., & Haviland, M. G. (2013). Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of Personality Assessment*, *95*, 129-140.

Reise, S. R., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In S. P. Reise & D. A. Revicki (Eds.), *Handbook of item response theory modeling: Applications to typical performance assessment* (pp. 13-40). New York, NY: Routledge.

Rhemtulla, M., Brosseau-Liard, P. E., & Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, *17*, 354-373.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137-150.

Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016b). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment*, *98*, 223-237.

Samejima, F. (1977). Weakly parallel tests in latent trait theory with some criticism of classical test theory. *Psychometrika*, *42*, 193-198.

Stucky, B. D., Thissen, D., & Orlando Edelen, M. (2013). Using logistic approximations of marginal trace lines to develop short assessments. *Applied Psychological Measurement*, *37*, 41-57.

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

Thurstone, L. L. (1935). *The vectors of mind*. Chicago, IL: University of Chicago Press.

Vigil-Colet, A., Lorenzo-Seva, U., & Morales-Vives, F. (2015). The effects of ageing on self-reported aggression measures are partly explained by response bias. *Psicothema*, *27*, 209-215.

Yuan, K. H., Chan, W., Marcoulides, G. A., & Bentler, P. M. (2016). Assessing structural equation models by equivalence testing with adjusted fit indexes. *Structural Equation Modeling*, *23*, 319-330.

Yule, G. U. (1907). On the theory of correlation for any number of variables, treated by a new system of notation. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *79*, 182-193.