Article

On the Added Value of Multiple Factor Score Estimates in Essentially Unidimensional Models

Educational and Psychological Measurement
2019, Vol. 79(2) 249–271

DOI: 10.1177/0013164418773851
journals.sagepub.com/home/epm
SAGE

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

Abstract

Measures initially designed to be single-trait often yield data that are compatible with both an essentially unidimensional factor-analysis (FA) solution and a correlated-factors solution. For these cases, this article proposes an approach aimed at providing information for deciding which of the two solutions is the most appropriate and useful. The procedures we propose are an FA extension of the "added-value" procedures initially proposed for subscale scores in educational testing. The basic principle is that the multiple FA solution is defensible when the factor score estimates of the primary factors are better measures of these factors than score estimates derived from a unidimensional or second-order solution. Methodologically, new results are obtained, and relations with factor indeterminacy measures and second-order FA are discussed. The procedures have been implemented in a noncommercial and widely known program for exploratory FA. The functioning of the proposal is assessed by means of a simulation study, and its usefulness is illustrated with a real-data example in the personality domain.

Keywords

factor score estimates, marginal reliability, added-value principle, second-order factor analysis

Applications of the factor-analysis (FA) model to item analysis and individual scoring conventionally use a random-regressors two-stage approach (e.g., McDonald,

[1]Universitat Rovira i Virgili, Tarragona, Spain

Corresponding Author:
Pere J. Ferrando, Facultad de Psicología, Universidad "Rovira i Virgili", Carretera Valls s/n, Tarragona 43007, Spain.
Email: perejoan.ferrando@urv.cat

1982). In the first stage (calibration), the structural (item) parameters are estimated using a limited-information procedure. In the second stage (scoring), the structural estimates are taken as fixed and known, and factor score estimates are obtained for each respondent.

A literature review shows that it is the first stage that receives by far the most attention (e.g., Fabrigar, Wegener, MacCallum, & Strahan, 1999). However, many instruments that are developed or calibrated with FA were designed for individual assessment purposes, and when this is the case, it is the scoring stage that should be the most important (Cliff, 1977; McDonald, 2011). More specifically, for this type of instrument, our position is that the ultimate aim of the FA application is to provide determinate, accurate, and valid factor score estimates for each potential respondent from the population for which the test is intended.

Of the various FA models that can be used in applications, the unidimensional (Spearman) model provides the clearest and most univocal interpretation of how this instrument functions at the calibration stage (McDonald, 1982, 2011). Furthermore, if the number of items is large enough, this model also allows most of the individuals in the population to be accurately measured. Most instruments that were designed to be single-trait, however, fail to meet the strict requirements for unidimensionality (mainly uncorrelated residuals) of Spearman's model (Furnham, 1990; Reise, Bonifay, & Haviland, 2013), and when this occurs, a multiple correlated-factors solution is usually fitted to the data. Indeed, in goodness-of-fit terms, the more parameterized multiple model will always fit better than Spearman's, and if enough factors are specified, a well-fitting solution is likely to be obtained. The choice of the most appropriate model, however, is a complex issue that goes beyond pure model-data fit.

As well as bad model-data fit, several other problems are expected to arise when a unidimensional model is fitted to data that are clearly multidimensional (Ferrando & Lorenzo-Seva, 2010; Reise et al., 2013). An initial problem is differentially biased item parameter estimates. A second is loss of information that might have implications in validity studies and in assessment. Finally, a third problem is factor score estimates that lack univocal interpretation and which reflect the impact of multiple sources of variance. On the other hand, however, fitting multidimensional models to data that are essentially unidimensional is likely to lead to an endless spiral of alternative solutions, make theory unnecessarily complex, and give rise to weak factors of little substantive interest (Ferrando & Lorenzo-Seva, 2017; Furnham, 1990; Reise et al., 2013). These weak factors, in turn, may yield factor score estimates that are indeterminate and unreliable, and which cannot provide accurate individual measurement (Ferrando & Navarro-González, 2018; Beauducel, Harms, & Hilger, 2016).

There are several procedures and indices for assessing the quality of psychometric FA solutions (Ferrando & Lorenzo-Seva, 2018; Rodriguez, Reise, & Haviland, 2016a, 2016b), and they should be routinely used to prevent the potential problems above. As Ferrando and Navarro-González (2018) and Reise et al. (2013) noted, however, there are many applications in which (a) an FA solution can be considered

to be essentially unidimensional, (b) the multiple solution is clear and interpretable, and (c) both the unidimensional and multidimensional solutions attain the standards of strength, replicability, and accuracy. And, when these conditions arise, it is hard to determine which solution will be the most appropriate and useful. Further information for deciding on this issue can be gained by comparing the factor score estimates provided by both solutions to see which of them most accurately measures the dimensions which were established at the calibration stage. This is the idea behind the present proposal.

In the context of educational measurement, and using classical test theory (CTT), Haberman (2008) considered the counterintuitive situation in which subscale scores are no better indicators of the subscale construct than total test scores, and argued that, to be considered useful, subscores must provide more accurate measurements of their construct than total scores. When this requirement is met, subscores are said to have ''added value.'' Statistical procedures based on these principles have been proposed and discussed in Haberman (2008), Haberman and Sinharay (2013), Sinharay (2010, 2013), Sinharay and Haberman (2008), and Sinharay, Puhan, and Haberman, (2011) in both CTT and item response theory (IRT) frameworks.

The ''added value'' principles and procedures mentioned above can be used in the FA context considered in this article to help choose the most appropriate FA model. The basic idea is to assess the extent to which factor score estimates in a primary factor derived from a multiple correlated solution are more accurate predictors of the corresponding ''true'' factor scores than factor score estimates derived from a single general factor. When they are, the correlated model is expected to provide more nontrivial information than can be obtained from the unidimensional model. However, if the unidimensional-based score estimates are as good as or better than the primary estimates at predicting the primary factors, then the choice of the multiple model is not justified.

Methodologically, the present article proposes procedures for assessing the ''added value'' principles on factor score estimates. So, our proposal is an adaptation of an existing principle in a different field. This type of adaptation appears to be new, and so are some of the results. Furthermore, relations between the FA-based added-value principles and issues such as factor indeterminacy and second-order FA solutions are established and discussed. This treatment also seems to be new. Overall, we believe that our proposal is a potentially useful addition to the factor analytic toolbox and allows the researcher to go beyond relying only on goodness-of-fit statistics. As far as novelty is concerned, we note that several authors have already recommended that the quality of the factor score estimates be routinely assessed and reported in those FA applications in which they are relevant (Ferrando & Lorenzo-Seva, 2018; Grice, 2001; Rodriguez et al., 2016a, 2016b). However, for the scenario considered here, this recommendation is only a necessary first step because, if the factor score estimates do not attain acceptable quality standards, the present proposal is no longer necessary.

Basic Results

We shall consider that the FA model is fitted by using the two-stage approach described above. In the first stage we shall assume that item calibration is based on the interitem correlation matrix but no other restrictions are imposed. So, calibration might be based on (a) any type of unrestricted or restricted FA solution and (b) both the linear model and the nonlinear model for ordered categorical variables (e.g., Muthén, 1984), which we shall denote here by CVM-FA (categorical variable methodology-FA). In the first case, calibration is based on the Pearson interitem correlation matrix. In the second case, it is based on tetrachoric (binary responses) or polychoric (graded responses) matrices. Tetrachoric-based FA can be viewed as an alternative parameterization of the IRT multidimensional two-parameter normal-ogive model, whereas polychoric-based FA can be viewed as an alternative parameterization of Samejima's (1969) multidimensional normal-ogive graded response model (see Ferrando & Lorenzo-Seva, 2013).

We turn now to the scoring stage, in which we shall use the terminology *true factor scores* to refer to the latent factor scores in the model (McDonald & Burr, 1967) and *factor score estimates* to refer to the corresponding predictors. This terminology makes it possible to establish clear relations with the CTT principles used in previous developments. However, its use should be qualified. Strictly speaking, because of the factor indeterminacy problem described below, there are no ''true'' parameters to be approximated by the estimates, but rather an infinity of random variables that are ''criterially'' latent factors. So, the term *estimates* is not strictly correct in the usual statistical sense either (see e.g., Maraun, 1996).

Let $\hat{\theta}_{ik}$ be the factor score estimate of individual i in the k factor, and let θ_{ik} be the corresponding true factor score. As in Samejima (1977), we can write

$$\hat{\theta}_{ik} = \theta_{ik} + \varepsilon_{ik}, \tag{1}$$

where ε_{ik} denotes the measurement error. As is usual in single-group FA, we shall assume that θ_k is distributed with zero expectation and unit variance. We shall further assume that: (a) $\hat{\theta}_{ik}$is conditionally unbiased (i.e., $E(\hat{\theta}_{ik}|\theta_{ik}) = \theta_{ik}$), and (b) the conditional distribution of $\hat{\theta}_k$ for fixed θ_k is normal. If (a) is fulfilled, then it follows that $E(\hat{\varepsilon}_{ik}|\theta_{ik}) = 0$, so the measurement errors are uncorrelated with the true trait levels. It then follows that the squared correlation between $\hat{\theta}_k$ and θ_k is

$$\rho^2_{(\hat{\theta}_k\theta_k)} = \frac{Var(\theta_k)}{Var(\hat{\theta}_k)} = \frac{1}{1+Var(\varepsilon_k)} = \frac{1}{1+E(Var(\varepsilon_{ik}|\theta_{ik}))} = \rho_{(\hat{\theta}_k\hat{\theta}_k)} \tag{2}$$

So (2) is both (a) the squared correlation between the true trait levels and their corresponding estimates and (b) the ratio of true variance to observed variance. Both are the standard definitions of a reliability coefficient (Lord & Novick, 1968), so (2) is taken here as the reliability of the factor score estimates. This type of reliability was initially proposed by Green, Bock, Humphreys, Linn, and Reckase (1984), who denoted it by *marginal reliability*, a term that we shall also use here. Provided that

$\text{Var}(\hat{\varepsilon}_{ik}|\theta_{ik})$ remains relatively uniform across trait levels, the coefficient in (2) is representative of the overall precision of the scores as measures of the corresponding factor (Brown & Croudace, 2015).

Expression (2) can be interpreted not only as an overall index of precision but also as a factor determinacy index (see, e.g., Ferrando & Lorenzo-Seva, 2018). Because more than one set of factor score estimates that are compatible with a given structure obtained at the calibration stage can be constructed, factor score estimates are said to be indeterminate (e.g., Beauducel, 2011). Furthermore, in practical applications, the degree of indeterminacy is usually quantified by means of the correlation (or squared correlation) between the factor score estimates and the true factor scores they predict (e.g., Beauducel, 2011; Grice, 2001), which is indeed index (2). So, a high marginal reliability value also means that (a) the factor score estimates are good proxies for representing the true factor scores and (b) the different factor score estimates that are compatible with the calibration results are also highly correlated with one another (Guttman, 1955). This alternative conceptualization is useful for interpreting the results below.

We shall now denote by θ_{ig} the true factor scores in the single or general common factor and by θ_{ik} the true factor scores of the k primary factor in the multiple oblique solution. The general factor can be estimated in one of these two ways: (a) as the common factor obtained by fitting the unidimensional (Spearman) model to the inter-item correlation matrix or (b) as a second-order factor based on the correlation among the primary factors of the oblique solution. In this second way, the general factor is understood as a higher-order attribute shared by the primary factors.

While choice (b) above is theoretically defensible, fitting Spearman's model to data which can strictly be considered to be multidimensional is more questionable. However, the scenario considered here assumes that the data are close enough to unidimensionality that no bias is expected in the loading estimates of the one-factor solution (Ferrando & Lorenzo-Seva, 2018; Rodriguez et al., 2016a, 2016b). When this is the case, the one-factor solution is a close approximation to the second-order solution (e.g., Mulaik & Quartetti, 1997). More specifically, if the second-order model is correct, Spearman's model becomes a particular case of it in which both unique and error variances are taken as residual variances (Rindskopf & Rose, 1988), and the one-factor loadings are given by

$$\lambda_j = \sum_k \lambda_{jk}\alpha_{kg} \tag{3}$$

where λ_{jk} is the loading of the j item on the k primary factor, and α_{kg} is the loading of the k primary factor on the g second-order factor.

The Added Value of Multiple Factor Score Estimates

Assume first that the true factor scores in the k primary factor are to be linearly predicted from the factor score estimates on the general factor

$$\theta'_{ik} = \beta_{gk}\hat{\theta}_{ig}. \tag{4}$$

From standard regression theory and the results above, it follows that

$$\beta_{gk} = \frac{\rho(\theta_k, \hat{\theta}_g)}{\sqrt{1 + Var(\varepsilon_g)}}. \tag{5}$$

And the mean squared error of prediction (MSE) is

$$MSE(\theta'_k - \theta_k) = 1 - \rho^2(\theta_k, \hat{\theta}_g). \tag{6}$$

By recalling that the factors are scaled to have zero mean and unit variance, it follows that the proportional reduction in MSE (PRMSE) when using the general factor score estimates rather than the mean of θ_k is simply $\rho^2(\theta_k, \hat{\theta}_g)$.

Assume next that the true factor scores on the k factor are now linearly predicted from the factor score estimates on the k factor itself. The prediction now becomes a variant of Kelley's (1947) formula and takes the form

$$\theta'_{ik} = \rho^2(\theta_k, \hat{\theta}_k)\hat{\theta}_{ik}, \tag{7}$$

which, by using the marginal reliability definition in (2), becomes

$$\theta'_{ik} = \rho_{(\hat{\theta}_k\hat{\theta}_k)}\hat{\theta}_{ik}. \tag{8}$$

The corresponding MSE is

$$MSE(Kelley) = 1 - \rho^2(\theta_k, \hat{\theta}_k) = 1 - \rho_{(\hat{\theta}_k\hat{\theta}_k)}, \tag{9}$$

and the PRMSE when the k factor score estimates are used instead of the mean of θ_k is simply $\rho_{(\hat{\theta}_k\hat{\theta}_k)}$.

In the present context, Haberman's (2008) rationale is that, for the k factor score estimates to have added value, the PRMSE based on these estimates must be greater than the PRMSE based on the general factor score estimates. Given the results above, Haberman's criterion in the present context implies that:

$$\rho_{(\hat{\theta}_k\hat{\theta}_k)} \geq \rho^2(\theta_k, \hat{\theta}_g) \tag{10}$$

must be fulfilled for considering that the k factor score estimates have added value.

Criterion (10) is only of theoretical interest because the correlation between the true factor scores in k and the estimated factor scores in the general factor is not known. To provide operational criteria, we shall now consider the two ways in which the general factor can be estimated.

Consider first that θ_g is obtained by fitting Spearman's model to the appropriate interitem correlation matrix. Next, factor score estimates derived from this model can be obtained in the scoring stage. For each individual, both the factor score estimates

on the general factor and the factor score estimates derived from the oblique model are then available, so the product moment correlations $r(\hat{\theta}_k, \hat{\theta}_g)$ can be empirically computed. By standard attenuation theory (e.g., Lord & Novick, 1968) and the basic results above, it follows that Criterion (10) can be operationalized as

$$\rho^2_{(\hat{\theta}_k \hat{\theta}_k)} \geq r^2(\hat{\theta}_k, \hat{\theta}_g). \tag{11}$$

We shall denote (11) as the empirical criterion because it is based on the empirical product–moment correlation computed from the corresponding set of factor score estimates in the sample.

We turn now to the second way of obtaining the general factor. The structural equation for the second-order FA model can be written as

$$\theta_{ik} = \alpha_{kg}\theta_{ig} + \xi_i. \tag{12}$$

In the present scaling, the second-order loading α_{kg} is also the product–moment correlation between the primary factor k and the general factor. If follows then that criterion (10) can be written as

$$\frac{\rho_{(\hat{\theta}_k \hat{\theta}_k)}}{\rho_{(\hat{\theta}_g \hat{\theta}_g)}} \geq \alpha^2_{kg}. \tag{13}$$

We shall denote Criterion (13) as a model-based criterion because the correlation between the true factor scores in k and g is now estimated as a structural parameter when the second-order FA model is fitted.

Expression (13) clearly shows that the determinants of the added value in the present context are (a) the correlation between the true factor scores in k and g, (b) the reliability of the general factor score estimates, and (c) the reliability of the primary factor score estimates. So, when the general factor is strongly related to the primary factors, and the reliability of its score estimates is larger than the primary reliability estimates, no added value will be obtained and this leads to choosing the unidimensional or second-order solution. Note that these conditions imply that the primary factors are highly correlated with each other and that the general factors are stronger and less indeterminate than the primary factors. On the contrary, if the primary factor score estimates are highly reliable and determinate, and the relation between them and the general factor score estimates is only moderate, the expected result will be added-value, and this will justify choosing the multiple model. This functioning agrees with the previous mechanisms obtained in the CTT context (Haberman, 2008; Sinharay et al., 2011).

Further Extensions: Weighted Averages

An additional question in the schema discussed above is whether the combined use of the factor score estimates in the primary factor and the factor score estimates in the general factor can substantially increase the precision with which the true primary

factor scores can be predicted with respect to each of the predictors individually. In practice, the most interesting scenario is when the primary factor score estimates have added value, in which case the question to be asked is whether the PRMSE attained can be even further improved by also considering the general factor score estimates as predictors. In a more general context, the basic idea described so far was considered by Wainer et al., (2001) and is known as subscore augmentation: augmenting data from a given subscale by using information from other parts of the test (or total test scores in our framework). The weighted averages approach we shall propose in this section is an FA adaptation of a proposal by Haberman (2008), which, in turn, can be considered to be a particular case of subscore augmentation (Haberman & Sinharay, 2010).

Assume that the true factor scores in the k factor are to be linearly predicted from both (a) the factor score estimates in the k factor and (b) the factor score estimates in the general factor

$$\theta'_{ik} = \beta_{kk}\hat{\theta}_{ik} + \beta_{gk}\hat{\theta}_{ig}. \tag{14}$$

By using standard results in multiple correlation analysis together with results in the previous sections, an empirical estimate of the PRMSE when using (14) can be written as

$$R^2_{\theta_k(\hat{\theta}_k,\hat{\theta}_g)} = \rho_{(\hat{\theta}_k\hat{\theta}_k)} + \omega, \tag{15}$$

where

$$\omega = \frac{\rho^2(\theta_k,\hat{\theta}_g)\left[\frac{1}{\rho_{(\hat{\theta}_k\hat{\theta}_k)}} + \rho_{(\hat{\theta}_k\hat{\theta}_k)} - 2\right]}{1-\rho^2(\theta_k,\hat{\theta}_g)}. \tag{16}$$

And where $\omega \geq 0$. So, the PRMSE when using (14) is always equal to or greater than the proportional reduction solely based on the k factor score estimates $\rho_{(\hat{\theta}_k\hat{\theta}_k)}$. It is equal (i.e., $\omega = 0$) when $\rho_{(\hat{\theta}_k\hat{\theta}_k)} = 1$, which makes sense: when the primary factor score estimates are perfectly reliable and determined, the proportional reduction is 1, and the accuracy cannot be improved. Inspection of (16) also shows that the proportional reduction increases as $\rho_{(\hat{\theta}_k\hat{\theta}_k)}$ decreases and $\rho^2(\theta_k,\hat{\theta}_g)$ increases, which again makes sense: the less reliable the primary factor score estimates are, and the more related the primary factor is to the general factor, the more information can be borrowed from the general factor score estimates.

The corresponding model-based result derived from fitting the second-order FA model (12) is again (15) with

$$\omega = \frac{\alpha_{kg}{}^2\rho_{(\hat{\theta}_k\hat{\theta}_k)}\rho_{(\hat{\theta}_g\hat{\theta}_g)}\left[\frac{1}{\rho_{(\hat{\theta}_k\hat{\theta}_k)}} + \rho_{(\hat{\theta}_k\hat{\theta}_k)} - 2\right]}{1-\alpha_{kg}{}^2\rho_{(\hat{\theta}_k\hat{\theta}_k)}\rho_{(\hat{\theta}_g\hat{\theta}_g)}}. \tag{17}$$

Previous CTT-based proposals have stated that the PRMSE in (15) based on both predictors should be substantially larger than each of the PRMSEs individually for considering that the augmented subscores (14) have practical utility (Haberman & Sinharay, 2010). However, how much larger it has to be also depends on the purposes for which subscores are used. Haberman and Sinharay (2013) proposed a more objective criterion that can tentatively be used in the present proposal. To simplify notation, denote by $PRMSE_2$ the proportional reduction (15) obtained with both predictors, and $PRMSE_1$ the best reduction obtained with a single predictor. So when both predictors are used precision is substantially larger than when a single predictor is used if

$$PRMSE_2 - PRMSE_1 > 0.1(1 - PRMSE_1). \tag{18}$$

That is to say, when $PRMSE_2$ reduces the distance of $PRMSE_1$ from 1 by at least 10%.

Factor Score Estimates and Reliability Estimates

The procedures described in the sections above assume that the factor score estimates are unbiased, which implies that (a) the true factor scores and the measurement errors are uncorrelated, so (b) the variance of the factor score estimates is the sum of the ''true'' and ''error'' components (see Equation 2). Estimators in common use such as maximum likelihood (ML), Bayes modal (MAP), or Bayes expected a posteriori (EAP) are asymptotically unbiased (asymptotically in this context means that the number of items increases without bound). For finite item sets, however, this is not generally the case.

We start by considering applications based on the linear FA model. In this case the ML estimates obtained under normality assumptions are Bartlett (1937) factor score estimates, whereas Bayes estimates (both MAP and EAP) are the regression factor score estimates for the oblique model (Thurstone, 1935), both of which can be obtained in closed form. Bartlett score estimates are conditionally unbiased even for finite item sets (e.g., McDonald, 2011). However, regression estimates are inwardly biased (i.e., regressed toward the mean) and their variance is smaller than the unit variance of the true factor scores they predict (Krijnen, Wansbeek, & Ten Berge, 1996). Finally, in the linear model the conditional error variance in (2) does not depend on θ_k so for both ML and Bayes score estimates the marginal reliability can be obtained in closed form (see, e.g., Ferrando & Lorenzo-Seva, 2018). This estimated reliability is generally higher for the regression (Bayes) factor score estimates (Beauducel et al., 2016). Overall, and given the results described so far, the procedures proposed in this article are expected to be correct in the linear case when based on Bartlett-ML estimated factor scores.

We turn now to the CVM-FA–based applications. In this case, neither the ML estimates nor the Bayes estimates are unbiased in finite item sets (Lord, 1986): ML factor

scores estimates are outwardly biased, whereas MAP and EAP estimates are inwardly biased.

The biases described above potentially affect both the correlations between factor score estimates and (in some cases) the marginal reliability estimates, so they are expected to have an impact on the comparisons and criteria proposed here. The relevance of this impact, however, depends on several factors. In general, it is expected to be greater in short tests and in empirical comparisons (11). In model-based comparisons, only the marginal reliability estimate is expected to be affected in some cases.

We start by considering ML estimates. The main problem here is that, in CVM-FA, finite estimates do not exist for extreme patterns, and implausibly large estimates might be obtained at both ends of the true θ range (Lord, 1986). However, if the estimates are constrained to be within a reasonable range (say -4 to $+4$), the outward bias is not expected to be a great problem in practice. Furthermore, corrections that make ML scores even less biased in finite item sets, such as Warm's (1989) WLE, can also be used.

With ML estimates, the marginal reliability (2) can be estimated either empirically or on the basis of the information function (e.g., Brown & Croudace, 2015). In the first case the empirical variance of the estimated factor scores is used in the denominator of (2). In the second case, the expectation of the information can be obtained by using quadrature approximations (see Ferrando, Navarro-González, & Lorenzo-Seva, 2017, for details). Next, the marginal reliability estimate is computed as

$$\rho_{(\hat{\theta}_k \hat{\theta}_k)} = \frac{1}{1 + E(\frac{1}{I(\theta_{ik})})}, \tag{19}$$

where $I(\theta_i)$ is the amount of information at the i trait level. While the empirical marginal estimate can be affected by the outward bias of the ML estimates (i.e., increased variance of the score estimates) estimate (19) is, in principle, free from bias.

The discussion so far indicates that, overall, ML estimates are more in agreement with the basic measurement Equation (1) than Bayes estimates. So, the inward bias and shrunken variance of these estimates (especially EAP) will be more problematic for the present proposals also in the CVM case. However, CVM-FA–based Bayes scoring provides finite and plausible estimates for all response patterns (Bock & Mislevy, 1982), and mainly for this reason, this is the most common type of scoring implemented in programs that perform CVM-FA. So, the use of MAP and EAP scores in the procedures proposed here must be addressed.

In principle, the marginal reliability of Bayes estimates can be obtained by using the squared posterior standard deviation (PSD) values as if they were error variances in Equation (2)

$$\hat{\rho}_{(\hat{\theta}_k \hat{\theta}_k)} = \frac{1}{1 + E(PSD^2(\hat{\theta}_{ik}))}. \tag{20}$$

As the number of items increases, the posterior distribution approaches normality (Chang & Stout, 1993) and the PSD becomes equivalent to an asymptotic standard error (Bock & Mislevy, 1982). So, for sets of more than, say, 10 items, the reliability estimate (20) is expected to be reasonably correct. For very short item sets, however, the PSDs are smaller than the standard errors because of the additional information contributed by the prior. So, a correction in this case can be obtained from the following approximate relation, which holds when the distribution of θ is standard normal (Wainer & Mislevy, 2000):

$$PSD(\hat{\theta}) \cong \frac{1}{\sqrt{I(\hat{\theta}) + 1}}. \tag{21}$$

The correction consists of transforming the PSDs into information amounts using (21) and then using the information values obtained in (19) to estimate the marginal reliabilities. When used together with the model-based criterion (13) this reliability correction is expected to lead to essentially correct results.

In the empirical-criterion case (11), however, the correlation $r(\hat{\theta}_k, \hat{\theta}_g)$ when based on Bayes (especially EAP) score estimates is expected to be upwardly biased with respect to the value that would be expected if the score estimates were conditionally unbiased. For this case, we propose the following simple correction. Denote by $s(\hat{\theta}_{kB})$and $s(\hat{\theta}_{gB})$ the observed standard deviations of the Bayes score estimates, and define the corrected standard deviations as

$$s_c(\hat{\theta}_{kB}) = \sqrt{1 + E(PSD^2(\hat{\theta}_{ik}))}$$

$$s_c(\hat{\theta}_{gB}) = \sqrt{1 + E(PSD^2(\hat{\theta}_{ig}))}. \tag{22}$$

The corrected empirical correlation $r_c(\hat{\theta}_k, \hat{\theta}_g)$ now becomes

$$r_c(\hat{\theta}_k, \hat{\theta}_g) = r(\hat{\theta}_k, \hat{\theta}_g) \frac{s(\hat{\theta}_{kB}) s(\hat{\theta}_{gB})}{s_c(\hat{\theta}_{kB}) s_c(\hat{\theta}_{gB})}. \tag{23}$$

Implementation

The proposals made here have been implemented and tested in an experimental version of FACTOR (Lorenzo-Seva & Ferrando, 2013), a well-known, free exploratory factor analysis program. They are now available at http://psico.fcep.urv.cat/utilitats/factor/ in the 10.08.01 release of the program. In order for them to be reported in the outcomes, users must select the option *Assess the added value of multiple factor score estimates* on the menu *Configure advanced indices related to the factor model*. Users must also decide whether the *Empirical* or the *Model based* approaches are to be computed. The outcomes are printed under the section *Added value of multiple factor score estimates*.

Simulation Study

An initial simulation study was undertaken to assess the sensitivity of the decision mechanisms to their main determinants: (a) reliability of the factor score estimates and (b) magnitude of the interfactor correlations. Furthermore, given the initial nature of the study, we considered only the simplest and most basic scenario, characterized by (a) only two primary factors, (b) continuous variables and Bartlett's ML score estimates, and (c) the empirical approach. This last choice is indeed the only one possible given that a second-order factor cannot be obtained on the sole basis of two primary factors.

Independent Variables

The study was based on a full $3 \times 3 \times 3 \times 7$ design with a total of 189 conditions and 200 replicas per condition. The independent variables were (1) sample size N = 200, 400, 800; (2) number of indicators m = 10, 20, 30; (3) loading value sizes: low (0.3), medium (0.5), and high (0.7); and (4) interfactor correlation: ranging from 0.20 to 0.80 in increments of 0.10. Variable (1) is expected to introduce more or less random sampling error in the different conditions, whereas variables (2), (3), and (4) are the main theoretical determinants of the outcomes of the procedure. Note specifically that the reliability of the factor score estimates largely depends on the number of the items that define the factor and the magnitude of their loadings (see, e.g., Ferrando & Lorenzo-Seva, 2018).

In all cases, the simulated patterns consisted of a bidimensional independent-clusters solution, with $m/2$ items defining each factor, and with the same loading value for all the items. Thus, for example, in the condition m = 10, with a low loading value, the pattern matrix would have the form

$$\mathbf{P} = \begin{bmatrix} .3 & 0 \\ .3 & 0 \\ .3 & 0 \\ .3 & 0 \\ .3 & 0 \\ 0 & .3 \\ 0 & .3 \\ 0 & .3 \\ 0 & .3 \\ 0 & .3 \end{bmatrix}$$

Dependent Variables

In each condition, the dependent variables were the PRMSE based on the general factor score estimates and the PRMSE based on the primary factor score estimates. To simplify the results the PRMSEs corresponding to the first and second factors were averaged, thus providing a single value per condition.

Results

The outcomes of the simulation study clearly showed that the impact of the sample size was virtually negligible. For this reason, we decided to present only the results averaged across sample sizes (the full results are available from the authors). They are shown in Figure 1. The thick solid line is the average PRMSE based on the general factor, whereas the dotted line is the average reduction based on the primary factors.

Overall, results in Figure 1 are meaningful and add interesting information. To start with, the PRMSE based on the primary factors is simply the reliability of the primary factor score estimates, which, in this case, does not depend on the interfactor correlation values (so it is a horizontal line in the graph). It does, however, clearly depend on the number of indicators and the magnitude of the loadings, as it should (note how the dotted line rises as a function of these determinants).

In the low-loading conditions, the outcomes of the procedure would lead to the unidimensional model being chosen in all the conditions considered. This result makes sense: with low loadings the primary factors are unreliable and poorly defined and the twice-as-long general factor leads to better predictions even when interfactor correlations are low.

In the medium-loading conditions the results are more complex, and the outcomes vary mainly as a function of test length and interfactor correlation. Note that with $m = 20$ the bidimensional model would only be chosen when the interfactor correlation is very low, whereas with $m = 30$ the unidimensional model would be the model of choice when the interfactor correlation exceeds the 0.50 threshold.

Finally, in the high-loading conditions, the results still make sense. When, in addition to the high loading, the number of indicators per factor is also relatively high, the primary factors are well defined and the derived scores are reliable. In these cases, the bidimensional model would be the model of choice except when the interfactor correlation is very high (i.e., when the simulated model is virtually unidimensional).

As a final comment, it is also worth noticing the compensatory way in which the number of items and the magnitude of the loadings act in determining the reliability of the primary factor score estimates, as the first graph in each row of Figure 1 is very similar to the last graph in the previous row.

An Illustrative Example

The Smoking Habits Questionnaire (SHQ) is a 22-item measure developed to assess different situations that stimulate the desire to smoke in habitual smokers. It was designed to measure three primary dimensions: (a) stress or relief from stress (9 items), (b) activity (7 items), and (c) boredom (6 items). However, it was also considered suitable for use as a general measure that assesses the desire to smoke across a variety of situations. The SHQ items ask the subjects to imagine themselves in a given situation and to rate on a 5-point scale their desire to smoke. The standard

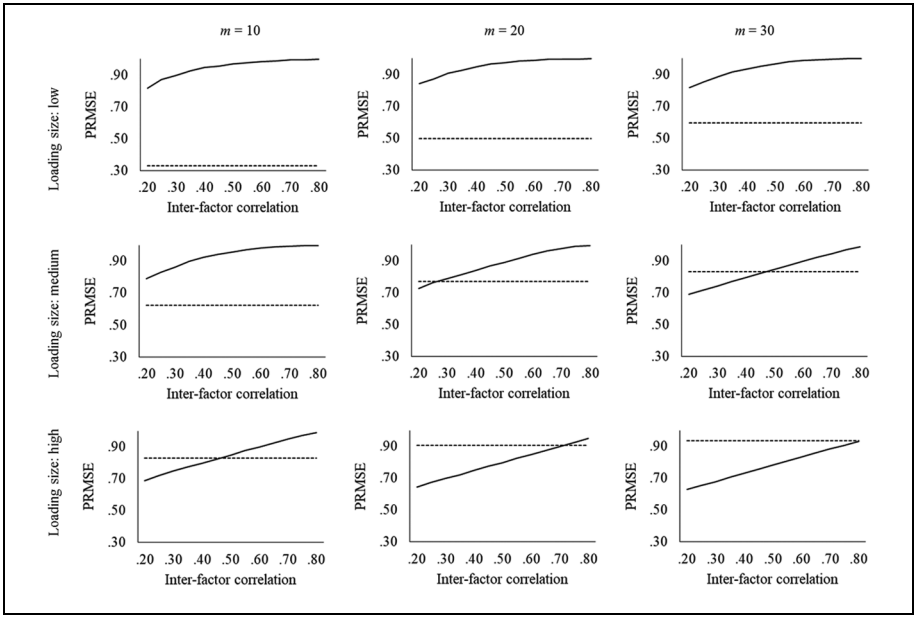

Figure 1. Results of the simulation study

analyses (e.g., Ferrando & Lorenzo-Seva, 2000) yield a clear solution in three substantially correlated factors that closely agrees with the structure expected a priori.

In this example, we reanalyzed the SHQ data used in Ferrando and Lorenzo-Seva (2000), which was based on a sample of 255 participants. For illustrative purposes, both the linear and the CVM FA models were fitted to this data, and in both cases, calibration was carried out by using robust unweighted least squares estimation as implemented in FACTOR. First, Spearman's model was fitted to the appropriate interitem correlation matrix. Second, an unrestricted oblique solution in three factors was obtained by using Promin (Lorenzo-Seva, 1999) rotation. Finally, a second-order solution with a single general factor was obtained based on the primary interfactor correlation matrix. Because only three primary factors were specified, the second-order solution is just-identified, so the fit is the same as that of the oblique solution.

Goodness of model-data fit was assessed by using both the conventional approach and the recent proposal by Yuan, Chan, Marcoulides, and Bentler (2016) based on equivalence testing. This is a new approach to assessing the fit of structural equation models that aim to endorse a model under the null hypothesis rather than reject it. However, adjusted cutoff values derived from this perspective (*T*-size values) are derived at present only for the root mean square error of approximation (RMSEA) and the comparative fit index (CFI) measures based on the linear model. In our analyses, both measures were based on the second-order (mean and variance) corrected chi-square statistic proposed by Asparouhov and Muthen (2010). Model-data fit

results are in Table 1. Because they are very similar for the linear and CMV FAs, they will be discussed together.

Overall, the fit of Spearman's models does not reach the limits of acceptability, whereas the fit of the tridimensional model is excellent by all standards. Regarding equivalence testing in particular, we note that the minimum tolerable sizes of model misspecification (*T*-sizes) for both RMSEA and CFI are very good in this model. However, the explained common variance index indicates that 83% to 84% of the common variance in the SHQ items can be explained by a single general factor. This value is above the cutoff values proposed in the literature (Ferrando & Lorenzo-Seva, 2018; Rodriguez et al., 2016a, 2016b), so it supports using the SHQ as a general measure.

For each FA model (linear and explained common variance), Table 2 shows the unidimensional pattern and the Promin rotated pattern with the dominant loadings boldfaced. The bottom of the table shows the Burt–Tuker congruence coefficients for assessing the similarity of the solutions. This is a measure of profile similarity (see Lorenzo-Seva & Ten Berge, 2006) that is defined as

$$\varphi(x,y) = \frac{\sum x_i y_i}{\sqrt{\sum x_i^2 \sum y_i^2}}. \tag{24}$$

Expression (24) was used to assess the congruence between the corresponding columns of the loading matrices. The overall congruence was obtained as the average of the column congruences.

As Table 2 shows, the congruence values are in all cases high enough to consider the two solutions virtually identical (Lorenzo-Seva & Ten Berge, 2006). So, we will again discuss them together. First, Spearman's solution exhibits positive manifold with substantial loadings for all the items, and Hancock and Mueller's (2001) *H* index is very high, suggesting that the single factor is strong, well defined, and replicable (Ferrando & Lorenzo-Seva, 2018; Rodriguez et al., 2016a, 2016b). Second, the solution in three factors is quite clear and is close to the structure expected ''a priori.'' Furthermore, the generalized *H* indices (see Ferrando & Lorenzo-Seva, 2018) are acceptably high in all cases, suggesting that all the three primary factors are strong, well-defined, and replicable. Overall, the results in Tables 1 and 2 make it hard to decide on what the most appropriate modelling is in this case.

Tables 3 (linear FA) and 4 (CVM FA) show the results of the procedures proposed in this article. The factor score estimates were Bartlett-ML in the linear case and EAP in the CVM case. In the latter case, both the marginal reliability estimates and the empirical correlations were corrected as described in Equations (21) to (23). The marginal reliability estimates of the general factor were 0.93 (linear) and 0.97 (CVM) when based on Spearman's model, and 0.92 (linear) and 0.96 (CVM) when based on the second-order model.

There is substantial agreement among the results obtained across procedures (empirical and model-based) and models (linear and CVM). So, the common results will be discussed first. As for the main determinants: (a) the three primary factors are

Table 1. Goodness-of-Fit Results for the Illustrative Example.

(a) Unidimensional solution

	RMSEA	95% CI RMSEA	*T*-size RMSEA	CFI	*T*-size CFI	GFI	*Z*-RMSR	ECV
Linear-FA	.086	(.073; .090)	.10 (mediocre)	.95	.93 (close)	.96	.090	.83
CVM-FA	.086	(.068; .089)		.97		.96	.100	.84

(b) Tridimensional solution

	RMSEA	95% CI RMSEA	*T*-size RMSEA	CFI	*T*-size CFI	GFI	*Z*-RMSR
Linear-FA	.011	(.010; .012)	.015 (close)	.99	.97 (excellent)	.99	.043
CVM-FA	.032	(.010; .033)		.99		.99	.048

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; ECV = explained common variance; FA = factor analysis; CVM = categorical variable methodology.

Table 2. Factor Solutions for the Illustrative Example.

	Linear solution				CVM solution			
Items	GF	PF1	PF2	PF3	GF	PF1	PF2	PF3
i1	.491	**.635**	−.138	.021	.575	**.660**	−.106	.066
i2	.617	**.646**	.078	−.059	.701	**.643**	.179	−.053
i3	.637	**.728**	.054	−.099	.776	**.729**	.198	−.077
i4	.575	**.749**	−.099	−.043	.638	**.772**	−.043	−.045
i5	.650	**.595**	−.057	.173	.728	**.628**	−.038	.214
i6	.610	**.705**	.104	−.157	.657	**.757**	.139	−.188
i7	.590	**.589**	.012	.037	.648	**.627**	.042	.037
i8	.698	**.534**	.099	.138	.746	**.590**	.125	.109
i9	.614	**.483**	−.007	.203	.663	**.523**	.027	.184
i10	.509	.093	−.206	**.735**	.549	.092	−.229	**.800**
i11	.684	.037	.143	**.645**	.731	.047	.147	**.679**
i12	.608	−.022	.186	**.572**	.647	–.046	.214	**.612**
i13	.604	−.074	.075	**.748**	.674	−.097	.108	**.817**
i14	.548	.120	−.234	**.784**	.602	.170	−.274	**.828**
i15	.649	−.120	.384	**.529**	.679	−.146	.407	**.567**
i16	.530	.097	−.148	**.693**	.576	.101	−.168	**.758**
i17	.651	.150	**.653**	−.041	.701	.162	**.703**	−.042
i18	.656	**.558**	.143	.018	.712	**.559**	.214	.015
i19	.636	.032	**.851**	−.116	.675	.019	**.902**	−.105
i20	.661	−.015	**.706**	.103	.705	−.020	**.756**	.111
i21	.633	.037	**.844**	−.119	.679	.085	**.861**	−.134
i22	.619	−.103	**.499**	.356	.651	−.102	**.524**	.365
GH-index	.932	.898	.897	.889	.949	.919	.922	.917

Note. CVM = categorical variable methodology. Factor Congruence: GF = 0.9994; PF1 = 0.9988; PF2 = 0.9934; PF3 = 0.9990; Overall = 0.9976.

substantially correlated, so the correlations between the general factor score estimates and the primary factor score estimates are also high; (b) the marginal reliabilities of the primary factor score estimates are rather high; and (c) the marginal reliability of the general factor score estimates is higher than that of any of the primary estimates. Results (a) and (c) above would run against added-value conclusions. However, the high marginal reliabilities (and so degree of determinacy) of the primary factors tips the balance in favor of the added value decision. Note that this decision is less clear in the case of the first primary factor (stress) which is the one that is most related to the general factor. In fact, in one of the cell results (empirical approach, linear FA) no added value is considered. Finally, we note that the present added-value results are in agreement with the results of the simulation study discussed above, as the conditions of the empirical study are similar to those in the first two graphs of the last row in in Figure 1 with an interfactor-correlation value around 0.50.

As for the common weighted-averages results shown at the bottom of Tables 3 and 4, according to criterion (18), when both the general factor score estimates and

Table 3. Added-Value Results for the Illustrative Example. Linear FA.

(a) Interfactor correlation matrix and basic estimates

	F1	F2	F3	$\rho_{\hat{k}\hat{g}}$	α_{kg}	$\rho_{\hat{k}\hat{k}}$
F1	1			0.90	0.95	0.88
F2	0.64	1		0.81	0.77	0.88
F3	0.56	0.46	1	0.78	0.70	0.88

(b) Proportional MSE reduction

	Empirical		Model based	
	From $\hat{g}$	From $\hat{k}$	From $\hat{g}$	From $\hat{k}$
F1	0.92	0.88	0.83	0.88
F2	0.76	0.88	0.54	0.88
F3	0.71	0.88	0.45	0.88

(c) Proportional MSE reduction. Weighted averages

	Empirical		Model based	
	From both $\hat{g}$ and $\hat{k}$	From $\hat{k}$	From both $\hat{g}$ and $\hat{k}$	From $\hat{k}$
F1	0.99	0.88*	0.96	0.88*
F2	0.93	0.88*	0.90	0.88*
F3	0.92	0.88*	0.89	0.88

Note. FA = factor analysis; MSE = mean squared error of prediction.
* = substantial PRMSE according to (18).

the primary factor estimates are used the PRMSE is substantially larger than the PRMSE when the best single predictor is used in virtually all cases (those marked with an asterisk).

In spite of the general agreement of the results discussed so far, consistent discrepancies also appear in Tables 3 and 4. With regard to the procedures, the general-factor PRMSEs are always larger when obtained empirically than when based on the second-order-model. And, as for the across-model comparisons, the marginal reliability estimates of all, general, and primary factor score estimates are always larger under the CVM-FA model. Together these results imply that the differences in PRMSEs that lead to the added-value decision are largest under model-based estimation and CVM-FA modelling. Also, given the smaller marginal reliabilities obtained in the linear case, it is under this model that the weighted-averages increases are largest.

The results obtained so far are now summarized. Although the SHQ can be justifiably used as a unidimensional measure, the three-factor solution is clear, strong, and interpretable. Furthermore, the factor score estimates derived from the multiple solution are more accurate measures of the corresponding dimension than the general

Table 4. Added-Value Results for the Illustrative Example. CVM FA.

(a) Interfactor correlation matrix and basic estimates

	F1	F2	F3	$\rho_{\hat{k}\hat{g}}$	α_{kg}	$\rho_{\hat{k}\hat{k}}$
F1	1			0.87	0.92	0.93
F2	0.70	1		0.78	0.76	0.93
F3	0.66	0.54	1	0.75	0.72	0.92

(b) Proportional MSE reduction

	Empirical		Model based	
	From $\hat{g}$	From $\hat{k}$	From $\hat{g}$	From $\hat{k}$
F1	0.81	0.93	0.83	0.93
F2	0.65	0.93	0.54	0.93
F3	0.60	0.92	0.45	0.92

(c) Proportional MSE reduction. Weighted averages

	Empirical		Model based	
	From both $\hat{g}$ and $\hat{k}$	From $\hat{k}$	From both $\hat{g}$ and $\hat{k}$	From $\hat{k}$
F1	0.96	0.93*	0.95	0.93*
F2	0.96	0.93*	0.94	0.93*
F3	0.95	0.92*	0.93	0.92*

Note. CVM = categorical variable methodology; FA = factor analysis; MSE = mean squared error of prediction.
* = substantial PRMSE according to (18).

factor score estimates based on a single-factor or a second-order solution. So, more relevant information for individual assessment is expected to be obtained if the tridimensional solution is chosen. Finally, if individual measurement accuracy is highly relevant (e.g., in high-stakes decisions), then the accuracy of the primary factor score estimates is expected to be substantially enhanced if factor score estimates based on the general factor are also used in the prediction. Note however that the use of this procedure implies either (a) fitting both the unidimensional and the multidimensional models or (b) fitting a second-order model.

Discussion

The approach proposed in this article has implications for two general issues in FA applications. The first one is that goodness-of-fit alone is not a sufficient criterion for deciding whether a given solution is appropriate or for choosing between two or more alternative solutions. The second is that the factor scoring results are the most relevant when the ultimate aim of the FA application is individual assessment. This

second issue does not appear to have received much attention in the FA literature (see Ferrando & Navarro-González, 2018). However, the first one has received some attention in recent years, in which awareness that overreliance on purely statistical model-data fit is not the way to go has clearly been growing (Ferrando & Lorenzo-Seva, 2018; Rodriguez et al., 2016a, 2016b). In agreement with both positions, we proposed an approach for deciding between two alternative FA solutions (unidimensional vs. correlated-multiple) that goes beyond model-data fit and is mostly based on the scoring results.

Methodologically, the present proposal is an adaptation to the FA context of proposals and principles that were originally made in the CTT context. Furthermore, the basic determinants we derived in the FA context are in agreement with those originally obtained in CTT (which is indeed a positive result). Even so, we believe that our FA adaptation is an original contribution, and as mentioned above, is of clear interest for the FA practitioner. It is simple, feasible, and provides an auxiliary source of information that allows the researcher to supplement decisions that, so far, are based solely on goodness-of-fit benchmarks. Furthermore, its implementation in a free, well-known, and user-friendly program makes it more likely to be used in practice.

It is acknowledged that the proposal has its share of limitations and points that deserve further study. For example, we found certain discrepancies across methods and FA models that deserve further, possibly simulation-based research. In particular, the PRMSEs are always larger when obtained empirically than when obtained via the second-order model, and the marginal reliability estimates based on the CVM-FA appear to be too optimistic.

The added-value criterion is an ''internal'' criterion, as are all the previous proposals on the accuracy and determinacy of the factor score estimates (e.g., Tucker, 1971). However, when (a) individual prediction is also an aim of the study and (b) relevant outside variables are available, then ''external'' criteria based on the relations between the estimated factor scores and these variables (e.g., Tucker, 1971) could also be considered to choose the most appropriate solution. Thus, a simple and immediate ''external'' extension of the approach proposed here would entail comparing the squared multiple correlation between the primary factor score estimates and the outside variable to the squared bivariate correlation between the general factor score estimates and this variable. Certain issues, however, such as the choice of the most appropriate type of factor score estimates in this context (e.g., Skrondal & Laake, 2001), the role of marginal reliabilities in correcting for attenuation effects, and the development of empirical and second-order model-based approaches clearly require further research.

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 project has been possible with the support of Ministerio de Economía, Industria y Competitividad, the Agencia Estatal de Investigación (AEI) and the European Regional Development Fund (ERDF) (PSI2017-82307-P).

References

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

Bartlett, M. S. (1937). The statistical conception of mental factors. *British Journal of Psychology*, *28*, 97-104.

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

Beauducel, A., Harms, C., & Hilger, N. (2016). Reliability estimates for three factor score estimators. *International Journal of Statistics and Probability*, *5*(6), 94-107.

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.

Chang, H., & Stout, W. (1993). The asymptotic posterior normality of the latent trait in an IRT model. *Psychometrika*, *58*, 37-52.

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

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*, 272-299.

Ferrando, P. J., & Lorenzo-Seva, U. (2000). Unrestricted versus restricted factor analysis of multidimensional test items: Some aspects of the problem and some suggestions. *Psicológica*, *21*, 301-323.

Ferrando, P. J., & Lorenzo-Seva, U. (2010). Acquiescence as a source of bias and model and person misfit: A theoretical and empirical analysis. *British Journal of Mathematical and Statistical Psychology*, *63*, 427-448.

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, Spain.

Ferrando, P. J., & Lorenzo-Seva, U. (2018). Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, *78*, 762-780. doi:10.1177/0013164417719308

Ferrando, P. J., & Navarro-González, D. (2018). Assessing the quality and usefulness of factor-analytic applications to personality measures: A study with the statistical anxiety scale. *Personality and Individual Differences*, *123*, 81-86. doi:10.1016/j.paid.2017.11.014

Ferrando, P. J., Navarro-González, D., & Lorenzo-Seva, U. (2017). Assessing the quality and effectiveness of the factor score estimates in psychometric factor-analytic applications. *Methodology* (Manuscript submitted for publication).

Furnham, A. (1990). The development of single trait personality theories. *Personality and Individual Differences*, *11*, 923-929.

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 theory1. *British Journal of Statistical Psychology*, *8*, 65-81.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204-229.

Haberman, S. J., & Sinharay, S. (2010). How Can Multidimensional Item Response Theory Be Used in Reporting of Subscores?. *ETS Research Report Series, RR-10-09 2010*(1).

Haberman, S. J., & Sinharay, S. (2013). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psychology*, *66*, 452-469.

Hancock, G. R., & Mueller, R. O. (2001). 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.

Kelley, T. L. (1947). *Fundamentals of statistics*. Cambridge, MA: Harvard University 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.

Lord, F. M. (1986), Maximum likelihood and Bayesian parameter estimation in item response theory. *Journal of Educational Measurement*, *23*, 157-162.

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

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

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57-64.

Lorenzo-Seva, U., & Ferrando, P. J. (2013). Factor 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and IRT models. *Applied Psychological Measurement*, *37*, 497-498.

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

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

McDonald, R. P. (2011). Measuring latent quantities. *Psychometrika*, *76*, 511-536.

McDonald, R. P., & Burr, E. J. (1967). A comparison of four methods of constructing factor scores. *Psychometrika*, *32*, 381-401.

Mulaik, S. A., & Quartetti, D. A. (1997). First order or higher order general factor? *Structural Equation Modeling*, *4*, 193-211.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*, 115-132.

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.

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51-67.

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. (1969). *Estimation of latent ability using a response pattern of graded scores* (*Psychometrika* Monograph No. 17). Iowa City, IA: Psychometric Society.

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

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150-174.

Sinharay, S. (2013). A note on assessing the added value of subscores. *Educational Measurement: Issues and Practice*, *32*, 38-42.

Sinharay, S., & Haberman, S. J. (2008). *Reporting subscores: A survey* (Research Memorandum 08-18). Princeton, NJ: ETS.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice*, *30*, 29-40.

Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, *66*, 563-575.

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

Tucker, L. R. (1971). Relations of factor score estimates to their use. *Psychometrika*, *36*, 427-436.

Wainer, H., & Mislevy, R. J. (2000). Item response theory, item calibration and proficiency estimation. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 61-101). Mahwah, NJ: LEA.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Rosa, K., Nelson, L., Swygert, K. A., & Thissen, D. (2001). Augmented scores—''Borrowing strength'' to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343-387). Mahwah, NJ: Erlbaum.

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*, 427-450.

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.