Article

Educational and Psychological Measurement
2019, Vol. 79(3) 437–461

DOI: 10.1177/0013164418824755
journals.sagepub.com/home/epm

An External Validity Approach for Assessing Essential Unidimensionality in Correlated-Factor Models

Pere Joan Ferrando[1] and Urbano Lorenzo-Seva[1]

Abstract

Many psychometric measures yield data that are compatible with (a) an essentially unidimensional factor analysis solution and (b) a correlated-factor solution. Deciding which of these structures is the most appropriate and useful is of considerable importance, and various procedures have been proposed to help in this decision. The only fully developed procedures available to date, however, are internal, and they use only the information contained in the item scores. In contrast, this article proposes an external auxiliary procedure in which primary factor scores and general factor scores are related to relevant external variables. Our proposal consists of two groups of procedures. The procedures in the first group (differential validity procedures) assess the extent to which the primary factor scores relate differentially to the external variables. Procedures in the second group (incremental validity procedures) assess the extent to which the primary factor scores yield predictive validity increments with respect to the single general factor scores. Both groups of procedures are based on a second-order structural model with latent variables from which new methodological results are obtained. 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, differential validity, incremental validity, second-order factor analysis

[1]Universitat Rovira i Virgili, Tarragona, Spain

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

Many psychological instruments, especially in the personality and attitude domains, were initially designed to be unidimensional or single trait (Furnham, 1990; Reise, Bonifay, & Haviland, 2013; Reise, Cook, & Moore, 2015). In most cases, however, failure of the item scores to meet the strict requirements of Spearman's factor analytic (FA) model lead to multiple FA solutions (generally exploratory) to be next fitted to the data. In turn, the results of these analyses lead to multiple correlated-factor solutions to be proposed as the most appropriate structure for many of these instruments (Ferrando & Lorenzo-Seva, 2018a, 2018b; Furnham, 1990; Reise et al., 2013; Reise et al., 2015). At the opposite extreme, many instruments that were originally designed to be multidimensional in fact yield data that are compatible with an essentially unidimensional solution (Floyd & Widaman, 1995; Reise et al., 2013; Reise et al., 2015).

The issue of deciding whether a unidimensional or a correlated-factor solution is the most appropriate for a measurement instrument is of the utmost practical importance. A unidimensional solution provides (a) the clearest and most univocal interpretation of how an instrument functions in the calibration stage (McDonald, 1982, 2011) and, generally, (b) the most accurate individual measurement in the scoring stage (Ferrando & Lorenzo-Seva, 2018a). However, forcing a unidimensional solution on data that are clearly multidimensional is likely to result in biased item parameter estimates, loss of information (this is one of the main points in the present article), and factor score estimates that cannot be univocally interpreted (see Ferrando & Lorenzo-Seva, 2018a). At the other extreme, ''splitting'' essentially unidimensional solutions into multiple solutions is expected to (a) lead to unnecessary complexities, (b) give rise to minor factors of little substantive interest, and (c) yield factor score estimates that are too unreliable and indeterminate (Beauducel, Harms, & Hilger, 2016; Ferrando & Lorenzo-Seva, 2018a; Furnham, 1990; Reise et al., 2013; Reise et al., 2015).

Standard goodness-of-fit (GOF) assessment of competing FA solutions is indeed a necessary first step in deciding what the most appropriate structure for the measure under study is. However, it is unlikely to provide a clear answer by itself. In pure GOF terms, the more parameterized multiple FA model will always fit better than Spearman's, and, if enough factors are specified, a well-fitting solution is likely to be obtained.

Recognizing that GOF alone is insufficient for judging the appropriateness and quality of an FA solution, several authors (Ferrando & Lorenzo-Seva, 2018a; Raykov & Marcoulides, 2018; Rodriguez, Reise, & Haviland, 2016a, 2016b) have proposed multifaceted approaches that go beyond pure model–data fit. This type of assessment, in turn, is expected to be particularly useful for deciding whether the most appropriate solution is unidimensional or correlated multiple. In general, the proposals to date focus on two broad groups of properties: (a) the strength, quality, and replicability of the FA solution and (b) the interpretability, accuracy, and determinacy of the factor score estimates derived from it. This second group of properties, which is what this article discusses, is particularly relevant to full psychometric applications, in which the ultimate aim is individual measurement in some form (e.g., assessment,

screening, classification, selection, or change). In this type of application, the accuracy and validity of the individual factor score estimates are the most important properties to be attained when fitting the FA model.

A common feature of all the approaches mentioned above is that they are ''internal'' in the sense that they only use the information provided by the item scores of the instrument under scrutiny. A literature review, however, shows that ''external'' or ''outside'' strategies have also been proposed for determining the most appropriate dimensionality. These strategies make use of the information contained in the validity relations between the score estimates and relevant external variables, whether these are objective criteria or scores on theoretically related measures. Relatively few ''external'' proposals have been made to date, they use different terminology, and they have different aims. In our opinion, however, most of them arise from two basic approaches that we shall refer to as differential and incremental. The basic question in differential proposals is whether the factor scores derived from the multiple primary factors relate differentially to the external variables (Carmines & Zeller, 1991; Floyd et al., 1992; Goldberg, 1972; Judge, Erez, Bono, & Thoresen, 2002). The basic question in incremental proposals is whether using primary scores in a multiple regression analysis leads to noticeable improvements in prediction with respect to using scores on a single general factor (Betts, Pickart, & Heistad, 2011; Coyle & Pillow, 2008; Floyd & Widaman, 1995; Mershon & Gorsuch, 1988). Overall, several authors (Floyd & Widaman, 1995; Goldberg, 1972; Mershon & Gorsuch, 1988) have considered that ''outside'' validity evidence is, in most cases, the ultimate criterion for judging the appropriateness of an FA solution.

This article proposes several procedures, based on an external validity approach, in which factor score estimates derived from a measurement FA model are related to theoretically relevant external variables. The proposed procedures are model based and aimed at assessing the two validity facets discussed above. They are particularly intended for those scenarios in which essentially unidimensional and correlated-factor solutions are plausible in ''internal'' terms, in the sense that in each competing solution, both the FA structure and the derived factor score estimates attain the standards of strength, replicability, and accuracy (see Ferrando & Lorenzo-Seva, 2018b; Reise et al., 2013). This scenario is quite common in practice and, when it occurs, the procedures we propose here can add information that can help reach a decision about the most appropriate dimensionality. As far as we know, our approach contains new results and developments, and we believe that it is a potentially useful addition when judging the appropriateness of FA solutions beyond pure model–data fit.

The procedures proposed in this article are based on a second-order factor model, which is extended to include validity relations with external variables. This model, which is described below in detail, is used as a basis for predicting how factor score estimates derived from the measurement submodel relate to the external variables when differential and incremental validity relations are present or not.

In the structural modeling framework, it has been more and more common practice to directly obtain validity estimates without having to obtain individual scores at

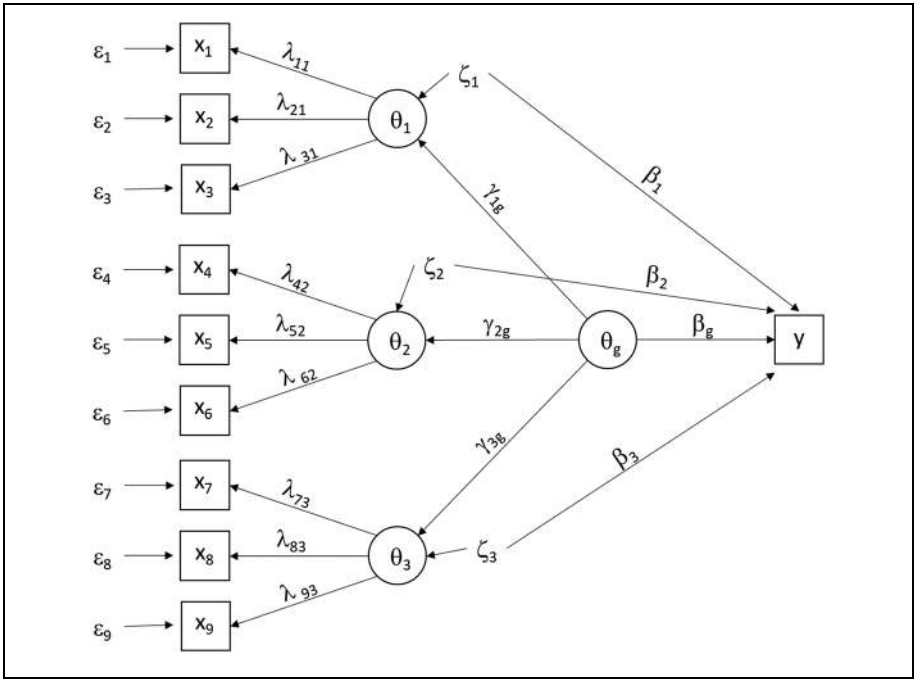

Figure 1. Path diagram for the basis model.

all (e.g., Curran, Cole, Bauer, Rothenberg, & Hussong, 2018). So some initial justification for the approach we have chosen is in order. First, as mentioned above, the FA solution whose dimensionality is to be assessed is expected to be used in a context in which individual measurement is the main aim of the application. Second, as described below, the full basis model is a nonstandard model that is not generally identified under the alternative solution, and furthermore, direct estimation becomes untractable or unfeasible in many applications found in practice (e.g., Gustafsson & Balke, 1993).

Rationale and Basic Results

Figure 1 shows the structural model that serves as a basis for our proposals. It is a second-order factor model that has been extended to include an external variable or criterion. The measurement part of the specific model depicted in the figure is a restricted solution with three indicators per primary factor, whereas the structural part considers three primary factors. Indeed, this representation is solely for purposes of simplicity. In our proposal, the measurement part of the model can be restricted or unrestricted (see Ferrando & Lorenzo-Seva, 2000) and the number of indicators and primary factors are not limited in any way. Furthermore, the depicted model

considers only a single external variable but, as in Carmines and Zeller (1991), multiple criteria can be considered. The scaling considered for the model is conventional, and the indicators x, the criterion y, and both the primary and the general factors are all scaled as standard variables, with zero mean and unit variance. Finally, the criterion is always taken to be a manifest variable.

The ''internal'' part of the second-order model in Figure 1 is conventional. The unique parts of the primary factors ζ_k are assumed to be uncorrelated between one another, which means that all the ''internal'' dependencies between the primary factors are accounted by for the general second-order factor. This assumption is reasonable if the unidimensional or the second-order FA models provide acceptable model–data fit results, which is the scenario considered here (e.g., Raykov & Marcoulides, 2018). Overall, the structural equations for the ''internal'' model are

$$x_{ij} = \lambda_{j1}\theta_{i1} + \cdots + \lambda_{jq}\theta_{iq} + \varepsilon_j \tag{1}$$

for the measurement submodel, and

$$\theta_{ik} = \gamma_{kg}\theta_{ig} + \xi_k \tag{2}$$

for the second-order latent structural model. Combining Equations (1) and (2) yields

$$x_{ij} = \left\lfloor \lambda_{j1}\gamma_{1g} + \cdots + \lambda_{jq}\gamma_{qg} \right\rfloor \theta_{ig} + \left\lfloor \lambda_{j1}\zeta_1 + \cdots + \lambda_{jq}\zeta_q + \varepsilon_j \right\rfloor = \alpha_j\theta_{ig} + \delta_j. \tag{3}$$

We turn now to the validity relations. In our basis model, each primary factor is related to the external variable y through two types of path: a direct path (β_k) and an indirect path (γ_{kg}; β_g). The direct paths model the potential validity relations between the unique parts of the primary factors and the criterion, a type of relation whose conceptual meaning has been discussed in the literature (e.g., Betts et al., 2011; Coyle & Pillow, 2008; Gustafsson & Balke, 1993; Judge et al., 2002; Nagy, Brunner, Lüdtke, & Greiff, 2017). Apart from random error, the unique parts of the primary factors may contain reliable specific variance that is not represented by the general factor but that is a structural part of the validity relations.

If all the direct paths are zero, it follows that all the validity relations between the primary factors and the external variables are mediated by the general factor, so no validity information will be lost if only the general factor is considered. This is the null model in our proposal. On the other hand, nonnegligible direct paths indicate that the unique parts of the primary factors are still related to the external variable beyond the relations that are mediated by the general factor. This is the alternative model in our proposal, and if it is correct, this additional validity information will be lost if a unidimensional solution is adopted.

As mentioned above, the alternative model considered here is not identified and so cannot be univocally estimated without additional identification constraints (see Nagy et al., 2017 for a related discussion). Furthermore, in most cases, it is a complex model likely to result in estimation problems. So it should be clear again that full structural estimation of this type of model is not an issue in the present proposal.

Rather, the estimation procedure we consider is three-stage procedure (e.g., Hoshino & Bentler, 2013). In the first stage (item calibration), the measurement FA model is fitted to the data. In the second stage (scoring), factor score estimates for each individual are obtained on the basis of the calibration results. In the third stage (external validity assessment), the factor score estimates are correlated to the relevant external variables or criteria. Because factor score estimates contain error and are biased, naive use of this three-stage procedure is expected to lead to biased (generally attenuated) validity estimates (e.g., Hoshino & Bentler, 2013). So, as discussed below, we shall propose in this article different procedures intended to address the inherent bias and error of the score estimates and provide correct validity inferences.

The corrected three-stage procedure described above is a good alternative when direct estimation of the full model is unfeasible, problematic, or too complex, but it also has other advantages (e.g., Curran et al., 2018; Devlieger & Rosseel, 2017; Hoshino & Bentler, 2013) of which we would like to mention two. First, in the case of binary and categorically ordered items, the resulting full model is usually a mixture of categorical and continuous (the criteria) variables, and the separate approach is preferable in most cases (Hoshino & Bentler, 2013). Second, in most validity studies (as in the one described below), the sample of individuals used for calibration and scoring purposes is larger than the subsample for which criteria or external measurements are available (Ghiselli, Campbell, & Zedeck, 1981). So more accurate calibration estimates can be obtained on the basis of the whole sample, while the third stage can then be performed on the subsample for which criteria measures are available.

We shall assume that item calibration is based on the interitem correlation matrix but no other restrictions are imposed. So it can be based on both the linear model and the nonlinear model for ordered categorical variables (e.g., Muthén, 1984), which we shall denote here by CVM (categorical variable methodology) FA (see Ferrando & Lorenzo-Seva, 2013, for a comparison of the linear and the CVM approaches in correlational terms).

A variety of approaches and solutions will be considered at the calibration stage. In the restricted FA case, a second-order solution can be directly fitted to the item scores. In the unrestricted FA case, the second-order factor loadings can be estimated by factoring the interfactor correlation matrix again. As for the unidimensional solution, the general factor can be taken to be the second-order factor (and so, understood as a higher order attribute shared by the primary factors), or obtained by directly fitting Spearman's model to the data. As far as this latter approach is concerned, we note that if the second-order solution is reasonably correct, the loadings on the one-factor solution will approach the second-order loadings. More specifically, the unique and residual variances will be taken as residual variances (see Equation 3), and the unidimensional loadings will be close to

$$\hat{\lambda}_j = \sum_k \lambda_{jk}\gamma_{kg}, \tag{4}$$

that is, the second-order loadings in Equation (3) (see, e.g., Mulaik & Quartetti, 1997; Rindskopf & Rose, 1988).

We turn now to the scoring and external validity stages. 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 (see Ferrando & Lorenzo-Seva, 2018a, for a further discussion). Let $\hat{\theta}_{ik}$ be the factor score estimate of individual i in the k primary factor, and let θ_{ik} be the corresponding true factor score. As in Samejima (1977) we can write,

$$\hat{\theta}_{ik} = \theta_{ik} + \varepsilon_{ik}, \tag{5}$$

where ε_{ik} denotes the measurement error. It is already assumed that θ_k is distributed with zero expectation and unit variance. Next, 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 $E(\varepsilon_{ik}|\theta_{ik}) = 0$, 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

$$\rho^2_{(\hat{\theta}_k\theta_k)} = \frac{\mathrm{Var}(\theta_k)}{\mathrm{Var}(\hat{\theta}_k)} = \frac{1}{1+\mathrm{Var}(\varepsilon_k)} = \frac{1}{1+E(\mathrm{Var}(\varepsilon_{ik}|\theta_{ik}))} = \rho_{(\hat{\theta}_k\hat{\theta}_k)}, \tag{6}$$

which is taken as the reliability of the factor score estimates (see Ferrando & Lorenzo-Seva, 2018b).

From Equation (5) and the assumptions above, it follows that the correlation between the primary factor score estimates in factor k and the relevant external variable y (i.e., the observed validity coefficient) is

$$\rho_{(\hat{\theta}_k y)} = \frac{\rho_{(\theta_k y)}}{\sqrt{\mathrm{Var}(\hat{\theta}_k)}}. \tag{7}$$

So

$$\hat{\rho}_{(\theta_k y)} = \frac{\rho_{(\hat{\theta}_k y)}}{\sqrt{\rho_{(\hat{\theta}_k\hat{\theta}_k)}}}. \tag{8}$$

Therefore, the disattenuated correlation between the estimated primary factor scores and the criterion is an unbiased estimate of the corresponding correlation between the true primary scores and the criterion (i.e., the true validity coefficient).

Differential Validity Assessment

Assume now that the disattenuated validity coefficients based on the primary factor score estimates have been obtained as in Equation (8). So, according to the basis model in Figure 1, the following relations would be expected

$$\begin{aligned} E(\hat{\rho}(\theta_1, y)) &= \gamma_{1g}\beta_g + \sqrt{1-\gamma^2}_{1g}\beta_1 \\ &\vdots \\ E(\hat{\rho}(\theta_k, y)) &= \gamma_{kg}\beta_g + \sqrt{1-\gamma^2}_{kg}\beta_k \\ &\vdots \\ E(\hat{\rho}(\theta_q, y)) &= \gamma_{qg}\beta_g + \sqrt{1-\gamma^2}_{qg}\beta_q \end{aligned} . \quad (9)$$

Under the null model in which the unique parts of the primary factors are uncorrelated with the criterion, the β_ks would all be zero, so the disattenuated coefficients should be proportional to the corresponding loadings of the primary factors on the second-order general factor. This result certainly makes sense because, as discussed above, all the validity relations in this case are mediated by the general factor. We note that the result so far discussed is a refinement of the proposal by Carmines and Zeller (1991), who stated that when there is no ''true'' multidimensionality in the validity sense, the primary factor scores should relate similarly to relevant criteria (p. 67). What we propose instead is that in this case they should relate to the criteria in the same proportion as how they relate to the general factor. So if the second-order loadings γ_{kg} are taken as fixed and known, the following result should hold if the null hypothesis is correct

$$\frac{\hat{\rho}(\theta_1, y)}{\gamma_{1g}} = \cdots = \frac{\hat{\rho}(\theta_k, y)}{\gamma_{kg}} = \cdots = \frac{\hat{\rho}(\theta_q, y)}{\gamma_{qg}} . \quad (10)$$

A simple procedure for testing the tenability of the null hypothesis of no differential validity would then be (a) obtain confidence intervals for the disattenuated validity coefficients by using Bootstrap resampling, (b) divide the endpoints of these intervals by the corresponding second-order loadings γ_{kg} (taken as fixed and known parameters), and (c) check whether the scaled intervals overlap (as they should under H_0) or not. If they overlap, we can conclude that the primary factor scores relate to the criterion as expected when all the validity relations are mediated by the general or second-order factor. In this case, the choice of the correlated-factor model would not provide additional validity information beyond that which can be obtained by using the unidimensional model.

Incremental Validity Assessment

The results in this section are more complex than those in the ''Differential Validity Assessment'' section, and for the sake of clarity and simplicity, we shall describe some of them using the simplest case of two primary factors. We shall also assume that all the validity relations (i.e., the β parameters in Figure 1) are positive. The loss of generality with these simplifications is assumed to be minor because (a) the generalization of the simplified results to a greater number of primary factors is straightforward and (b) the direction of the primary factors is usually arbitrary and can be reversed at one's convenience.

The basis approach proposed here is to compare (a) the correlation between the general or second-order factor score estimates and the criterion with (b) the multiple correlation between the primary factor score estimates and the criterion when both (a) and (b) are corrected for the measurement error in the primary factor score estimates. As shown below, when the null model holds, both correlations are expected to have the same value.

We shall first consider how the factor scores on the second-order or general factor are obtained. If the ''true'' factor scores on the primary factors were known, the second-order scores would be a weighted composite of the primary scores, and the weights would reflect the impact of the general factor on the corresponding primary factors. More specifically, under the assumption of conditional normality above, the maximum likelihood (ML) factor score estimates would be Bartlett's (1937) scores (see Ferrando & Lorenzo-Seva, 2018b, for a discussion). Now, if Bartlett's scores were obtained on the basis of the ''true'' primary factor scores, then the weights of the linear composite would be proportional to the signal-to-noise ratios $\gamma_{kg}/1-\gamma^2_{kg}$. These weights are ''optimal weights'' in the sense that they define the composite with maximal reliability (Penev & Raykov, 2006; Raykov, Gabler, & Dimitrov, 2016). In the specific case of factor scores, Ferrando (2008) labeled this composite as the ''maximal information'' composite.

Under the null model, and in the simplest case of two primary factors, the correlation between the second-order factor score estimates and the criterion would be

$$\begin{aligned} E(\rho_{\hat{\theta}_g y}) &= \beta_g \sqrt{\frac{v_1^2\gamma_1{}^2 + v_2^2\gamma_2{}^2 + 2v_1v_2\gamma_1\gamma_2}{v_1^2 + v_2^2 + 2v_1v_2\gamma_1\gamma_2}} \\ &= \beta_g saf(\hat{\theta}_g) \end{aligned} \tag{11}$$

where the *v*s are the weights of the composite, and *saf* is the ''structural attenuation factor'' term that reflects the structural errors of the primary factors (see Figure 1). Note from Equation (11) that *saf* will only reach unit value (i.e., no attenuation) when the γs are all 1 (i.e., when the primary factors are perfect indicators or markers of the second-order factor). Now, an estimate of $\rho_{\hat{\theta}_g y}$ denoted by $\hat{\rho}_{\hat{\theta}_g y}$ can be obtained by correcting the corresponding observed correlation (based on the primary factor score estimates) by the measurement error in these estimates. The expected value of this corrected correlation under H_0 is that given in Equation (11).

The relation (11), however, does not hold under the alternative model because in this case the corrected correlation above also contains an additional term that reflects the relations between the unique parts of the primary factors and the criterion. For the simplest case of two primary factors, the expected value of the corrected correlation under H_1 can be written as

$$E(\hat{\rho}_{\hat{\theta}_g y}) = \beta_g saf(\hat{\theta}_g) + \left[K\frac{\gamma_1}{\sqrt{1-\gamma_1^2}}\right]\beta_1 + \left[K\frac{\gamma_2}{\sqrt{1-\gamma_2^2}}\right]\beta_2. \tag{12}$$

where K is a constant term.

We turn now to the multiple regression results. Again, if the true primary factor scores are known, the prediction of the criterion from the primary factors would be a direct application of regression analysis. However, they are not, and fallible factor score estimates are used instead of the true factor scores. The measurement error in these primary estimates can be corrected via error-in-variables regression (e.g., Johnston, 1972, chap. 9). More specifically, Ferrando (2008) considered applying this type of regression to factor score estimates that fulfill the conditions discussed here. Let $\mathbf{S_P}$ be the covariance matrix between the estimated factor scores, and $\mathbf{S_{PT}}$ the corrected matrix with unit variances in the main diagonal (i.e., the estimated covariance matrix between the true factor scores). Next, let $\mathbf{s_{py}}$ be the vector containing the covariances between the estimated factor scores and the criterion. The error-in-variables estimated vector of weights is given by

$$\mathbf{w} = (\mathbf{S_{PT}})^{-1}\mathbf{s_{Py}}. \tag{13}$$

For the simplest case of two primary factors, let $[w_1, w_2]$ be the elements of $\mathbf{w}$. Then, an estimate of the multiple R that would be obtained if the true factor scores were known can be obtained as (see, e.g., McNemar, 1969, eq. 11.6)

$$R_c = \sqrt{w_1\hat{\rho}(\theta_1, y) + w_2\hat{\rho}(\theta_2, y)}. \tag{14}$$

From Equation (14), it can then be found that, under the null model, the expected value of R_c is

$$\begin{aligned} E(R_c) &= \beta_g \sqrt{\frac{w_1^2\gamma_1{}^2 + w_2^2\gamma_2{}^2 + 2w_1w_2\gamma_1\gamma_2}{w_1^2 + w_2^2 + 2w_1w_2\gamma_1\gamma_2}} \\ &= \beta_g saf(reg) \end{aligned}. \tag{15}$$

Now, if H_0 holds, the weights w in Equation (15) are proportional to the optimal weights v in Equation (11). This is because the optimal composite that maximizes reliability also maximizes validity for any criterion that is uncorrelated with the residual (structural in our case) errors (Penev & Raykov, 2006; Raykov et al., 2016). It then follows that the *saf* term is the same in both Equations (11) and (15), so the expected value of the corrected correlations to be compared is the same when the null model holds (i.e., when there is no incremental validity).

When the alternative model holds, the expected value of R_c becomes

$$E(R_c) = \beta_g saf(reg) + \left\lfloor Cw_1\sqrt{1-\gamma_1^2} \right\rfloor \beta_1 + \left\lfloor Cw_2\sqrt{1-\gamma_2^2} \right\rfloor \beta_2 . \tag{16}$$

If the expressions in Equations (12) and (16) are compared, it becomes clear that both are weighted averages of the β validity parameters, but their values are expected to differ. The general factor score composite aims to maximize the reliability or

information about the second-order factor, whereas the regression composite aims to maximize the prediction of the criterion. So, overall, under H_0, both Equations (12) and (16) are expected to have the same value, but under H_1, the expectation of Equation (16) is the maximal validity that can be obtained when correcting the primary factor score estimates for error, and so it is necessarily larger than the expectation of Equation (12).

As in the section above, a simple test for incremental validity is to obtain bootstrap confidence intervals for both R_c and $\hat{\rho}_{\hat{\theta}_g y}$ and check whether they overlap (as they should under H_0) or not. An alternative procedure, more in line with a related proposal by Raykov et al. (2016), is to compute the difference $R_c - \hat{\rho}_{\hat{\theta}_g y}$, obtain the Bootstrap confidence interval for this difference, and check whether the zero value falls within the interval or not. In a more general context, we note that the difference between two correlations (usually squared) is the most common operative measure of incremental validity (Haynes & Lench, 2003).

Some Additional Considerations

In a broader context, the method we propose here (a) separately estimates the measurement and structural (validity) parts of a complex model and (b) corrects the primary factor score estimates for error to attain unbiased estimates of the parameters of interest. In this context, then, our proposal is related to more general bias-correction proposals such as that of Croon's (2002), and Hoshino and Bentler's (2013). More specifically, in the differential case, our proposed correction is expected to provide unbiased estimates of the validity coefficients, whereas, in the incremental case, the common correction we propose provides a correct basis for comparing the two correlations of interest.

The two groups of procedures summarized above behave appropriately when the factor score estimates are conditionally unbiased, from which the classical requirements of error-in-variables regression (errors linearly independent of true levels with zero expectation and known reliability) are then obtained (see Ferrando, 2008). Strictly speaking, however, only Bartlett's ML estimates in the linear FA model are unbiased for finite item sets (e.g., McDonald, 2011). Regression or Bayes estimates are always inwardly biased, and, in the CVM FA model, all the factor score estimates in common use are biased to some extent. So, again strictly speaking, the results provided here must be considered as approximate in many cases.

Ferrando and Lorenzo-Seva (2018b) made a detailed discussion of the practical relevance of the problem above. As a summary, they found that ML estimates (the counterpart of Bartlett scores in the CVM case) also behaved quite well in the CVM FA model. As for regression (linear) and Bayes modal or expected a posteriori (CVM) scores, they proposed additional corrections mainly intended for the marginal reliability coefficients derived from the factor score estimates. Interested readers are referred to these corrections if they aim to implement the present proposal based on scores other than ML.

Sensitivity Determinants and a Simulation Study

In accordance with the developments above, two major factors are expected to determine the sensitivity with which differential and incremental validity will be detected when the alternative model is true. The first determinant is the relative strength of the weights of β_g and β_ks (see Figure 1). As the general factor becomes more strongly related to the criterion while the unique relations become weaker, the model increasingly approaches the null model so validity improvements are less likely to be detected.

The second factor concerns the relations between the vector of γ weights and the vector of β_k weights, and for clarity and simplicity, we shall consider that all these weights are positive. Now, all other things being constant, detection of differential and incremental validity is expected to be enhanced when the unique parts relate to the criterion in a different way to how the corresponding primary factors relate to the general factor. So sensitivity is expected to be high when the unique parts of the primary factors that are weakly related to the general factor are strongly related to the criterion, and vice versa (see Figure 1). On the contrary, when the profile of the γ and β_k vectors is similar (i.e., the unique parts of the primary factors most related to the general factor are also the most related to the criterion), detection is expected to be difficult. In more detail, if the γ and β_k vectors are proportional, then (a) the terms in Equation (10) are expected to be quite similar, as the differences will only depend on the residuals $\sqrt{1-\gamma_k^2}$ and (b) the corrected validity coefficients based on the general factor score estimates (Equation 12) and on the regression scores (Equation 16) are also expected to be similar, because the profile of weights on the maximum information, and the maximum validity composites approach one another. Indeed, if both determinants are taken into account, the detection of validity effects is most difficult when β_g is larger than the β_k*'s* and the profile of the γ and β_k vectors is similar.

The determiners discussed so far served as a basis for designing the simulation study described in this section. Given its preliminary nature, the study considered only the simplest setting based on continuous item scores, a single criterion, and Bartlett's ML score estimates. It consisted of two parts: (a) a preliminary study, in which the null model was correct in the population, and (b) the main study, based on different conditions under the alternative model. The preliminary study served mainly to check that the above predictions derived under H_0 were correct.

General Conditions

In all cases, pseudo-populations of 1,000,000 simulees were generated, and for each condition, 1,000 samples of the prescribed size were drawn from the corresponding pseudo-population. So, the study was based on 1,000 replicas per cell in all cases. Overall, 9,000 samples were assessed under H_0 and 54,000 under H_1.

Independent Variables

The H_0 study was based on a full 3×3 design. The independent variables were (a) sample size $N = 300, 500, 1{,}000$, and (b) number of primary factors $r = 3, 4, 5$. In all

conditions, the simulated patterns for the primary factors were independent-clusters solutions, with five items per factor and a constant loading of 0.8. Thus, for example, the pattern for the three-factor condition was

$$P = \begin{bmatrix} .8 & 0 & 0 \\ .8 & 0 & 0 \\ .8 & 0 & 0 \\ .8 & 0 & 0 \\ .8 & 0 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & .8 & 0 \\ 0 & 0 & .8 \\ 0 & 0 & .8 \\ 0 & 0 & .8 \\ 0 & 0 & .8 \\ 0 & 0 & .8 \end{bmatrix}$$

The remaining constant conditions were as follows. The β_g parameter for the general factor (see Figure 1) was 0.45, the βs for the unique factors were 0, and the γ loadings for the primary factors ranged from 0.30 to 0.8. The reliabilities of the primary factor scores were all 0.90, which follows from the settings of the design.

The main H_1 study was based on a full $3 \times 3 \times 3 \times 2$ design. The independent variables in this case were (a) sample size $N = 300, 500, 1{,}000$; (b) number of primary factors $r = 3, 4, 5$; (c) degree of agreement between the γ and β_k vectors; and (d) relative strength of the general and unique β parameters. With regard to this last independent variable, in the low general strength condition, the β for the general factor was 0.30, and the βs for the unique factors ranged from 0.20 to 0.60 while in the high general strength condition, the general β was fixed to 0.60 and the unique βs ranged from 0.10 to 0.40. As for the γ and β_k degree of agreement, the three levels were (a) disagreement, in which the γ and β_k values were set in the opposite order; (b) random ordering; and (c) agreement in which the order of the γ and β_k values was the same. Finally, the measurement or ''internal'' part of the model was the same under H_0 and H_1. So the internal constant conditions were the same as in the previous study.

Dependent Variables

In both studies, differential and incremental validity were assessed by using difference statistics. In the differential case, first, the scaled disattenuated validity coefficients in Equation (10) were computed, and next, their median value across the primary factors was obtained. In the H_0 study, a scaled value chosen at random was subtracted from the median value in each replication. In the H_1 study, the most

Table 1. Differential and Incremental Validities When H_0 Is True.

	Differential validity			Incremental validity		
	Mean	c5	c95	Mean	c5	c95
Overall	−0.002	−0.223	0.216	0.009	−0.008	0.032
r = 3	−0.003	−0.229	0.216	0.007	−0.008	0.025
r = 4	−0.004	−0.224	0.215	0.009	−0.008	0.031
r = 5	0.000	−0.214	0.221	0.012	−0.007	0.036
N = 300	−0.004	−0.261	0.255	0.014	−0.009	0.041
N = 500	−0.003	−0.228	0.225	0.009	−0.009	0.028
N = 1,000	0.000	−0.160	0.160	0.005	−0.007	0.017

Note. c5 = 5 percentile confidence interval; c95 = 95 percentile confidence interval.

extreme scaled value was subtracted from the median. This procedure provides a single difference statistic regardless of the number of factors. As for incremental validity, the chosen statistic was the $R_c - \hat{\rho}_{\hat{\theta}_g y}$ difference described above. In both cases, the results reported in the tables are the median values across the 1,000 replications as well as the interval defined by the 5 and 95 empirical percentiles.

Results

Table 1 shows the results of the H_0 study, which are quite clear: Neither spurious differential nor incremental validity are detected in any of the conditions (note that the zero value falls within the percentile confidence interval in all cases). Note also that the median value is always close to its expected zero value but that for both differential and incremental outcomes, it seems to get closer and closer to zero as the sample size increases.

The results for the more complex H_1 study are in Table 2 (differential) and Table 3 (incremental). Overall, they agree quite well with the theoretical expectations. For both types of validities, effects will be detected (boldfaced results in both tables) when the γ and β_k vectors, vectors disagree or are random, and the sensitivity will increase when the general weight is lower than the unique weights. As for differences, in the differential validity case, sensitivity also seems to increase with the number of factors and sample size, but not in the incremental validity case.

Neither differential nor incremental validity will be detected when the profiles of the γ and β_k vectors are in agreement. For differential effects, this result makes clear sense because the unique parts of the factors are essentially related to the criterion in the same way as the corresponding primary factors are related to the general factor (which is the main assumption under H_0). As for the lack of incremental effects in these conditions, the results suggest that the confounded (unique-general) validity effects have already been taken into account by the general factor score estimates, so that the multiple regression linear composite, with weights that are very similar to

Table 2. Differential Validities When H_0 is False.

Beta of general factor	Low (.30)			High (.60)		
Betas of primary factors	β–γ disagreement	Random	β–γ agreement	β–γ disagreement	Random	β–γ agreement
Overall	**1.012** (0.820; 1.242)	**1.07** (0.879; 1.269)	−0.009 (−0.253; 0.281)	**0.676** (0.498; 0.892)	**0.721** (0.536; 0.918)	0.008 (−0.234; 0.233)
r = 3	**0.939** (0.791; 1.131)	**1.04** (0.876; 1.201)	−0.016 (−0.278; 0.281)	**0.657** (0.494; 0.853)	**0.727** (0.554; 0.895)	0.017 (−0.196; 0.203)
r = 4	**0.994** (0.813; 1.201)	**1.048** (0.852; 1.240)	−0.063 (−0.250; 0.254)	**0.661** (0.482; 0.883)	**0.706** (0.523; 0.907)	0.012 (−0.241; 0.247)
r = 5	**1.104** (0.898; 1.305)	**1.124** (0.923; 1.322)	0.052 (−0.213; 0.301)	**0.712** (0.516; 0.929)	**0.729** (0.533; 0.943)	−0.005 (−0.256; 0.244)
N = 300	**0.994** (0.762; 1.279)	**1.068** (0.831; 1.312)	0.015 (−0.276; 0.315)	**0.651** (0.355 0.938)	**0.711** (0.483; 0.966)	0.014 (−0.267; 0.265)
N = 500	**1.015** (0.827; 1.240)	**1.072** (0.884; 1.267)	−0.004 (−0.253; 0.279)	**0.681** (0.512 0.887)	**0.724** (0.543; 0.914)	0.008 (−0.234; 0.228)
N = 1,000	**1.027** (0.874; 1.202)	**1.071** (0.932; 1.223)	−0.038 (−0.231; 0.231)	**0.698** (0.569 0.839)	**0.727** (0.596; 0.859)	0.003 (−0.183; 0.185)

Note. Centiles 5 and 95 are given in parentheses, and significant indices are printed in boldface.

those of the general factor score composite, does not add substantially to the prediction of the criterion. So, in these conditions, the outcomes of the proposed procedure would lead to the unidimensional model being chosen even when unique relations with the criterion do in fact exist. This lack of sensitivity may be considered a limitation of our proposal at the theoretical level, but not at the practical level. In effect, in this case, the choice of the unidimensional solution is justifiable in terms of validity and is more parsimonious than the multiple solution.

An Empirical Example

The Statistical Anxiety Scale (SAS; Vigil-Colet, Lorenzo-Seva, & Condon, 2008) is a narrow-bandwidth personality test intended to measure anxiety manifestations related to the encounter of statistics in any form or level. This type of manifestation is expected to hinder the learning process (Onwuegbuzie & Daley, 1999). So SAS scores are expected to be negatively related to academic performance, particularly to performance measures in which statistics is involved in any way. The SAS consists of 24 items and uses a 5-point graded response format.

Table 3. Incremental Validities When H_0 is False.

Beta of general factor	Low (.30)			High (.60)		
Betas of primary factors	β–γ disagreement	Random	β–γ agreement	β–γ disagreement	Random	β–γ agreement
Overall	**0.279** (0.195; 0.367)	**0.229** (0.127; 0.329)	0.020 (−0.012; 0.059)	**0.118** (0.057; 0.184)	**0.102** (0.043; 0.171)	0.009 (−0.016; 0.038)
r = 3	**0.270** (0.185; 0.356)	**0.170** (0.101; 0.245)	0.023 (−0.009; 0.062)	**0.108** (0.048; 0.171)	**0.077** (0.032; 0.127)	0.006 (−0.016; 0.031)
r = 4	**0.299** (0.216; 0.385)	**0.264** (0.185; 0.346)	0.021 (−0.013; 0.061)	**0.121** (0.060; 0.187)	**0.112** (0.053; 0.177)	0.009 (−0.016; 0.038)
r = 5	**0.268** (0.191; 0.349)	**0.253** (0.180; 0.330)	0.016 (−0.015; 0.053)	**0.125** (0.065; 0.192)	**0.116** (0.057; 0.183)	0.011 (−0.017; 0.044)
N = 300	**0.282** (0.176; 0.389)	**0.232** (0.115; 0.350)	0.024 (−0.017; 0.070)	**0.121** (0.044; 0.203)	**0.106** (0.034; 0.193)	0.013 (−0.020; 0.048)
N = 500	**0.279** (0.195; 0.364)	**0.228** (0.124; 0.327)	0.020 (−0.013; 0.056)	**0.118** (0.057; 0.181)	**0.101** (0.043; 0.168)	0.009 (−0.017; 0.036)
N = 1,000	**0.277** (0.216; 0.345)	**0.226** (0.135; 0.305)	0.017 (−0.007; 0.043)	**0.115** (0.071; 0.162)	**0.098** (0.051; 0.148)	0.005 (−0.013; 0.024)

Note. Centiles 5 and 95 are given in parentheses, and significant indices are printed in boldface.

The initial SAS study (Vigil-Colet et al., 2008), which was based on a sample of 159 undergraduates, obtained a clear three-factor structure, with factors labeled as ''asking for help anxiety'' (AHA), ''interpretation anxiety'' (IA), and ''examination anxiety'' (EA). This structure closely matched the intended structure when the measure was first designed. However, the factors were also found to be substantially and positively correlated with one another, and the authors found that the hypothesis of a general dimension of statistical anxiety running through the 24 SAS items was tenable. So they suggested that the SAS could be scored either as a tridimensional measure with separate scales or as a general scale.

The present example is an extended reanalysis of the data presented in Ferrando and Navarro-González (2018), which is based on a larger sample of 384 students. The general aim is to assess whether the correlated three-factor structure or the essentially unidimensional structure is more appropriate for SAS items. Given that the item responses are ordered categorical, the most appropriate FA model ''a priori'' is the CVM, which was the approach used in Ferrando and Navarro-González (2018). However, our preliminary analyses suggested that the results obtained with the linear FA model and the CVM FA model were virtually identical, a result that is not unusual

Table 4. Goodness-of-Fit Results for the Empirical Example.

	RMSEA	95% CI RMSEA	T-s RMSEA	CFI	T-s CFI	GFI	z-RMSR	ECV
1-Factor	.128	(.112; .139)	.14 (mediocre)	.90	.87 (mediocre)	.91	.090	.79
3-Factor	.032	(.021; .034)	.035 (close)	.99	.99 (excellent)	.99	.042	—

Note. RMSEA = root mean square error of approximation; CI = confidence interval; T-s RMSEA = T-size root mean square error of approximation; CFI = comparative fit index; T-s CFI = T-size comparative fit index; GFI = goodness-of-fit index; z-RMSR = root mean square of residuals; ECV = explained common variance (ECV measures closeness to unidimensionality).

in personality measures with five or more graded response items. So for the sake of brevity and clarity, only the simple linear FA-based results will be discussed here.

First-stage item calibration was carried out by using robust unweighted least squares estimation as implemented in the FACTOR program (Lorenzo-Seva & Ferrando, 2013). First, Spearman's model was fitted to the product-moment interitem correlation matrix. Second, an unrestricted oblique solution in three factors was obtained by using Promin rotation (Lorenzo-Seva, 1999). 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. Results of the item calibration were then taken as fixed and known and were used to obtain factor score estimates, which were Bartlett-ML scores, in the second step.

The most appropriate solutions were assessed in three stages: first, conventional GOF assessment was used to compare the fit of the two competing models; second, added-value assessment (Ferrando & Lorenzo-Seva, 2018b) was used to determine whether primary factor score estimates could be used instead of the general factor score estimates to nontrivially reduce the measurement error; and third, the external validity assessment proposed in this article.

Table 4 shows the goodness of model–data fit results obtained with the conventional approach and the recent proposal by Yuan, Chan, Marcoulides, and Bentler (2016) based on equivalence testing. All the measures that are dependent on the chi-square GOF statistic were based on the second-order (mean and variance) corrected chi-square statistic proposed by Asparouhov and Muthen (2010).

It seems clear that the fit of the unidimensional solution does not reach the limits of acceptability, whereas the fit of the tridimensional solution is excellent by all standards. As far as equivalence testing in particular is concerned, we note that the minimum tolerable sizes of model misspecification (T-sizes) for both RMSEA and CFI are very good in this case. At the same time, however, the explained common variance index indicates that 79% of the common variance in the SAS items can be explained by a single general factor (Ferrando & Lorenzo-Seva, 2018a; Rodriguez et al., 2016a, 2016b), which supports the proposal to use the scale as a general measure.

Table 5. Factor Solutions for the SAS Example.

Items	GF	F1	F2	F3
i1	.638	.109	.110	**.580**
i2	.371	−.171	**.765**	.009
i3	.675	**.880**	−.153	.015
i4	.575	.012	.053	**.664**
i5	.613	**.560**	.188	−.021
i6	.447	.023	**.760**	−.108
i7	.754	**.854**	−.002	.007
i8	.612	.227	**.334**	.218
i9	.613	.065	−.049	**.746**
i10	.325	−.108	**.666**	−.041
i11	.613	.104	−.045	**.695**
i12	.749	**.926**	−.095	.009
i13	.492	−.110	.037	**.713**
i14	.648	.181	.051	**.562**
i15	.532	−.174	−.014	**.892**
i16	.457	.120	**.274**	.196
i17	.758	**.937**	−.037	−.042
i18	.433	.061	**.540**	.011
i19	.561	.194	**.411**	.127
i20	.547	−.120	.024	**.811**
i21	.711	**.667**	.251	−.067
i22	.518	−.038	**.891**	−.052
i23	.696	**.901**	−.128	−.002
i24	.668	**.669**	.160	−.047
GH index	.936	.952	.887	.909

Note. SAS = Statistical Anxiety Scale; GF = general factor; GH index = generalized H index. Dominant loadings are printed in boldface.

The unidimensional and rotated three-factor solutions (with the dominant loadings boldfaced) are shown in Table 5. The multiple solution was virtually the same as that obtained in the initial SAS study, with the AHA (F1), IA (F2), and EA (F3) factors essentially defined by eight items each. The unidimensional solution has positive manifold with substantial loadings (greater than .30) 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, 2018a; Rodriguez et al., 2016a, 2016b). The solution in three factors, besides agreeing with theory, is quite clear and the generalized H (GH) indices (see Ferrando & Lorenzo-Seva, 2018a) are acceptably high in all cases, suggesting that all three primary factors are strong, well defined, and replicable. These results make it hard to decide on what the most appropriate solution for the SAS is.

We turn now to the added-value results based on the factor score estimates, which are provided in Table 6. It should be noted that (a) the three primary factors are substantially correlated, so the correlations between the general factor score estimates

Table 6. Internal Assessment: Added-Value Results for the Empirical Example.

Panel (a): Interfactor correlation matrix and basic estimates

	F1	F2	F3	$\rho_{\hat{k}\hat{g}}$	γ_{kg}	$\rho_{\hat{k}\hat{k}}$
F1	1			0.88	0.73	0.95
F2	0.45	1		0.64	0.61	0.89
F3	0.52	0.44	1	0.79	0.71	0.91

Panel (b): Proportional MSE reduction based on factor scores

	From $\hat{g}$	From $\hat{k}$
F1	0.81	0.95
F2	0.47	0.89
F3	0.70	0.91

Note. MSE = mean square error .

and the primary factor score estimates are also high, and (b) the marginal reliabilities of the primary factor score estimates are rather high. The estimated reliability of the second-order factor score estimates (i.e., the general factor) was 0.97, which is about the same as that of the first primary factor score estimates. The results in the lower panel (b) of Table 6 clearly suggest that, for the three primary factors, there is added value if the primary factor score estimates are used instead of the general factor score estimates.

So far, the GOF and the added-value results tend to favor the multiple solution as the most appropriate and informative. However, it will be interesting to see whether these conclusions, all of which are based on internal evidence, are maintained when the external validity procedures proposed in this article are considered.

As an appropriate external criterion for the SAS scores, we used the marks on a final statistical exam, which were available for 238 of the respondents. As expected from theory, the relations between the SAS scores and the criterion were all negative, and according to the conventions above, they were reversed to provide correlations that were always positive.

Table 7 provides the differential (panel a) and incremental (panel b) results based on the approach we propose. As for panel (a), the results suggest that the primary factors do not relate to the criterion in the way that should be expected from the null second-order model. Rather, the first factor (AHA) seems to be more strongly related and the second factor (IA) more weakly related to the criterion than could be predicted by their relations with the general factor.

With regard to panel (b), the results also appear to be clear: The prediction based on the primary score estimates is significantly more accurate than that based on the general factor score estimates when both correlations are corrected for measurement error. The overall conclusion, then, is that, although it is defensible to score the SAS

Table 7. External Validity Assessment.

Panel (a): Differential validity results

	$\widehat{\rho}_{\hat{\theta}_{ky}}/\gamma_k$	90% CI
F1	0.57	(0.44; 0.66)
F2	0.20	(0.14; 0.36)
F3	0.35	(0.21; 0.46)

Panel (b): Incremental validity results

$\widehat{\rho}_{\hat{\theta}_{gy}}$	90% CI	R_c	90% CI	dif	90% CI
0.32	(0.22; 0.41)	0.62	(0.51; 0.73)	0.30	(0.21; 0.40)

Note. CI = confidence interval; dif = differential.

as a unidimensional measure, accuracy and external information are lost if the unidimensional solution is chosen instead of the multiple solution in three correlated factors. This loss would be particularly important if, as assumed here, accurate individual measurement and/or prediction is the main focus of the application.

Discussion

The debate about how to determine the most appropriate dimensionality of FA solutions based on item responses is as old as the technique itself and has sometimes led to extreme positions. In particular, the development of FA as a ''proper'' statistical technique (e.g., Lawley & Maxwell, 1963) promised an objective and rigorous approach to this issue in the form of GOF testing. Particularly in recent decades, however, it has been clear that overreliance on GOF results is not the way to go and that any sound approach for assessing dimensionality must be multifaceted. In accordance with this view, some ''internal'' comprehensive approaches have recently been proposed (Ferrando & Lorenzo-Seva, 2018a; Raykov & Marcoulides, 2018; Rodriguez et al., 2016a, 2016b). In our opinion, they are a clear step forward toward arriving at parsimonious, clear, strong, and useful FA solutions in real applications.

The present article aims to expand the multifaceted internal approaches in existence to date by incorporating information about how the factor score estimates derived from competing plausible FA solutions relate to relevant external variables. Indeed, this type of outside or external approach has already been discussed in the literature (Betts et al., 2011; Carmines & Zeller, 1991; Coyle & Pillow, 2008; Floyd et al., 1992; Goldberg, 1972; Judge et al., 2002; Mershon & Gorsuch, 1988). At the methodological level, however, and to the best of our knowledge, the existing outside proposals are purely descriptive, and lack a clear methodological foundation. In contrast, ours is model based and allows differential and incremental validity to be more rigorously assessed.

At a more general level, our proposal is mostly intended for applications in which the ultimate aim of FA is individual measurement and external information is available. Contrary to the most common views (see Curran et al., 2018, for a discussion), we believe that the scoring stage is the most important part of FA in this case and, in accordance with this view, our proposal is mostly based on the scoring results.

Like any new proposal, this one has its share of limitations and points that deserve further study. Thus, the results of the simulation study are encouraging in general, but they are only preliminary, and further research of this type is still needed. More specifically, we believe that three main points require intensive research: (a) to ascertain the minimal degree of accuracy and determinacy of the factor score estimates that is required for obtaining correct validity inferences (see the discussion below), (b) to assess how the proposal behaves under the graded-response modeling, and (c) to assess the functioning of the proposal when regression (Bayes) factor score estimates are used instead of Bartlett's ML estimates. Point (c) includes also testing the bias-and-error corrections so far proposed for regression scores. Further applied research based on real data is also needed to clearly establish the practical usefulness of our proposal.

At a more applied level, the present proposal is mostly based on disattenuated and multiple correlations, both of which are prone to well-known empirical problems. With regard to the differential-validity indices, the disattenuated validity estimates are based, in turn, on reliability estimates, and if these are poor, the resulting indices can be misleading. This problem is expected to appear mainly when the factor score estimates are weak and unreliable. When this is the case, the indeterminacy of the factor score estimates is also high (Ferrando & Lorenzo-Seva, 2018a), which, in turn, implies that the external correlations with the criterion are not determinate and can vary over a wide range of values (Steiger, 1979). To sum up, the external correlations are unstable, and the correction for attenuation is expected to further increase this instability because the reliability estimates are themselves unstable. The potential problems summarized so far imply that our proposal must only be used for comparing competing solutions that lead to acceptable factor score estimates. This is indeed the recommendation we have been giving throughout the article. It makes little sense to use an additional auxiliary procedure if the previous internal procedures suggest that some of the solutions to be compared are unacceptable.

The most important potential problem in terms of the incremental validity results is the deflation of the multiple correlation coefficient when the regression equation is used in replication samples. This problem has been explicitly documented when the regressors are factor score estimates (Morris, 1979) and has been considered in previous external proposals (Betts et al., 2011; Goldberg, 1972; Haynes & Lench, 2003). When deflation occurs and is substantial, the assessment of incremental validity is indeed questionable. However, the relevance of this problem must be qualified. Previous research designs tended to use schemas with a large number of primary factors and samples that were not too large (e.g., Goldberg, 1972), a scenario which is indeed conducive to problems of shrunken *R*s. At the other extreme, the problem is

likely to be far smaller if the assessment is based on a large and representative sample, and the number of primary factors compared with the unidimensional solution is reasonably small (say 4 or 5 at most, which is the setting considered in the simulation study). Having said that, we acknowledge that a cross-validated design would reinforce the conclusions obtained in a single sample, but again, this practice would also be useful for the previous ''internal'' procedures. Finally, if applications of our proposal suggest that the problem is of practical importance, corrections for R_c could be considered in future developments.

Despite its limitations and the fact that further research is still needed, we believe that our proposal is of clear interest for practitioners who use FA for individual measurement purposes. It is simple, feasible, and provides an auxiliary source of information that enables decisions based on internal approaches to be supplemented. From a practical point of view, we note that our proposal can work with no need to use raw scores. So an *R* program, for example, with factor score, gamma estimates, and criterion scores as input would be quite easy to build.

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

Betts, J., Pickart, M., & Heistad, D. (2011). Investigating early literacy and numeracy: Exploring the utility of the bifactor model. *School Psychology Quarterly*, *26*, 97-107.

Carmines, E. G., & Zeller, R. A. (1991). *Reliability and validity assessment* (Vol. 17). Newbury Park, CA: Sage.

Coyle, T. R., & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence*, *36*, 719-729.

Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. A. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 195-223). Mahwah, NJ: Lawrence Erlbaum.

Curran, P. J., Cole, V. T., Bauer, D. J., Rothenberg, W. A., & Hussong, A. M. (2018). Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 860-875. doi:10.1080/10705511.2018.1473773

Devlieger, I., & Rosseel, Y. (2017). Factor score path analysis. *Methodology*, *13*, 31-38.

Ferrando, P. J. (2008). Maximizing the information and validity of a linear composite in the factor analysis model for continuous item responses. *Psicológica*, *29*, 189-203.

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. (2013). *Unrestricted item factor analysis and some relations with item response theory* (Technical Report). Tarragona, Spain: Universitat Rovira i Virgili, Department of Psychology.

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

Ferrando, P. J., & Lorenzo-Seva, U. (2018b, May 15). On the added value of multiple factor score estimates in essentially unidimensional models. *Educational and Psychological Measurement*, *79*, 249-271.

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.

Floyd, F. J., Haynes, S. N., Doll, E. R., Winemiller, D., Lemsky, C., Burgy, T. M., & . . . Heilman, N. (1992). Assessing retirement satisfaction and perceptions of retirement experiences. *Psychology and Aging*, *7*, 609-621.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286-299.

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

Ghiselli, E. E., Campbell, J. P., & Zedeck, S. (1981). *Measurement theory for the behavioral sciences: A series of books in psychology*. San Francisco. CA: W. H. Freeman.

Goldberg, L. R. (1972). Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs*, 72(2), 59.

Gustafsson, J. E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research*, *28*, 407-434.

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.

Haynes, S. N., & Lench, H. C. (2003). Incremental validity of new clinical assessment measures. *Psychological Assessment*, *15*, 456-466.

Hoshino, T., & Bentler, P. M. (2013). Bias in factor score regression and a simple solution. In A. R. de Leon & K. C. Chough (Eds.), *Analysis of mixed data* (pp. 43-61). Boca Raton, FL: Chapman & Hall/CRC Press.

Johnston, J. (1972). *Econometric methods*. New York, NY: McGraw-Hill.

Judge, T. A., Erez, A., Bono, J. E., & Thoresen, C. J. (2002). Are measures of self-esteem, neuroticism, locus of control, and generalized self-efficacy indicators of a common core construct? *Journal of Personality and Social Psychology*, *83*, 693-710.

Lawley, D. N., & Maxwell, A. E. (1963). *Factor analysis as statistical method*. London, England: Butterworth.

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

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.

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.

McNemar, Q. (1969). *Psychological statistics*. New York, NY: Wiley.

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

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

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, *55*, 675-680.

Morris, J. D. (1979). A comparison of regression prediction accuracy on several types of factor scores. *American Educational Research Journal*, *16*, 17-24.

Nagy, G., Brunner, M., Lüdtke, O., & Greiff, S. (2017). Extension procedures for confirmatory factor analysis. *Journal of Experimental Education*, *85*, 574-596.

Onwuegbuzie, A. J., & Daley, C. E. (1999). Perfectionism and statistics anxiety. *Personality and Individual Differences*, *26*, 1089-1102.

Penev, S., & Raykov, T. (2006). On the relationship between maximal reliability and maximal validity of linear composites. *Multivariate Behavioral Research*, *41*, 105-126.

Raykov, T., Gabler, S., & Dimitrov, D. M. (2016). Maximal criterion validity and scale criterion validity: A latent variable modeling approach for examining their difference. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 544-554.

Raykov, T., & Marcoulides, G. A. (2018). On studying common factor dominance and approximate unidimensionality in multicomponent measuring instruments with discrete items. *Educational and Psychological Measurement*, *78*, 504-516.

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. P., 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* (pp. 13-40). New York, NY: Routledge.

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

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

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

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

Steiger, J. H. (1979). The relationship between external variables and common factors. *Psychometrika*, *44*(1), 93-97.

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema*, *20*, 174-180.

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: A Multidisciplinary Journal*, *23*, 319-330.