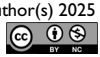
Original Research Article

Educational and Psychological Measurement
2025, Vol. 85(5) 857–881


sagepub.com/journals-permissions
DOI: 10.1177/00131644251319286
journals.sagepub.com/home/epm


Assessing the Properties and Functioning of Model-Based Sum Scores in Multidimensional Measures With Local Item Dependencies: A Comprehensive Proposal

Pere J. Ferrando[1], David Navarro-González[2] and Fabia Morales-Vives[1]

Abstract

A common problem in the assessment of noncognitive attributes is the presence of items with correlated residuals. Although most studies have focused on their effect at the structural level, they may also have an effect on the accuracy and effectiveness of the scores derived from extended factor analytic (FA) solutions which include correlated residuals. For this reason, several measures of reliability/factor saturation and information were developed in a previous study to assess this effect in sum scores derived from unidimensional measures based on both linear and nonlinear FA solutions. The current article extends these proposals to a second-order solution with a single general factor, and it also extends the added-value principle to the second-order scenario when local dependences are operating. Related to the added-value, a new coefficient is developed (an effect-size index and its confidence intervals). Overall, what is proposed allows first to assess the reliability and relative efficiency of the scores at both the subscale and total scale levels, and second, provides information on the appropriateness of using subscale scores to predict their own factor in comparison to the predictive capacity of the total score. All that is proposed is implemented in a freely available R program. Its usefulness is illustrated with an empirical example, which shows the distortions that correlated residuals may cause and how the various measures included in this proposal should be interpreted.

[1]Research Center for Behavior Assessment, Universitat Rovira i Virgili, Tarragona, Spain
[2]Universitat de Lleida, Lleida, Spain

Corresponding Author:
Fabia Morales-Vives, Departament de Psicologia, Facultat de Ciències de l'Educació i Psicologia, Universitat Rovira i Virgili, Carretera de Valls s/n, Tarragona 43007, Spain.
Email: fabia.morales@urv.cat

Keywords

sum scores, correlated residuals, relative efficiency, second-order solutions, added-value

This article deals jointly with three topics of great practical relevance in the noncognitive measurement domain. First, is the problem of related specificities or local dependences (LDs) among the items that form a measurement instrument. Second, in multidimensional instruments that can also be considered to measure a more general single dimension, is the convenience of using subscale scores instead of a single total score. The third topic, finally, is the interest or convenience of using simple sum scores instead of more informative and theoretically superior alternatives. The following brief, and necessarily limited, review of these topics will be made from the perspective of the general factor analytic (FA) framework (e.g., McDonald, 1985, 1999, 2000) adopted in this article.

The conventional FA-based modeling in psychometric applications is a random-regressor two-stage approach (McDonald, 1985), which, from a more general Structural Equation Modeling (SEM) view, corresponds to a limited-information estimation and fitting procedure (e.g., B. Muthén, 1984, 1993). The first stage is the structural stage itself and consists of fitting the prescribed FA solution to the inter-item covariance or correlation matrix. In psychometric terms, this first stage corresponds to the item calibration stage. Next, once an acceptable structural solution has been obtained, it is taken as fixed and known and used as a basis for obtaining individual score estimates at the second, scoring stage. The outcomes of the scoring stage include (or should include) not only the individual score point estimates but also measures of their accuracy and appropriateness (Ferrando & Lorenzo-Seva, 2019; Mansolf & Reise, 2018).

In FA, the first topic of related item specificities or local item dependencies has been mostly assessed at the bivariate level and is usually known under the terms ''correlated residuals'' or ''doublets'' (Mulaik, 2010). Regarding terminology first, in what follows we shall use all the terms so far introduced indistinctly, because, in the FA modeling, all refer essentially to the same phenomenon: That the pair of items share specific variance so that, once the influence of the common ''content'' factors has been partialled-out, they continue to covary due to a series of noncontent causes. This is a pervasive phenomenon in the noncognitive field, and the most relevant causes are: (a) model misspecification, (b) repeated presentation of identical items, (c) content or wording similarities, (d) similarities in the evoked situation, and (e) context effects (Bandalos, 2021; DeMars, 2020; Edwards et al., 2018; Ferrando & Morales-Vives, 2023; Lucke, 2005).

The ''traditional'' FA approach views the doublet phenomenon as a general problem of faulty test design, and the prescribed cure is to remove offending items until full local independence (LI) is attained (see, e.g., Ferrando & Morales-Vives, 2023 for a discussion). This position, however, is possibly too simplistic. While the

problem can be improved in most cases with a careful design (particularly when the causes are (a), (b), or (e) above), for constructs such as narrow-bandwidth normal-range traits or clinical traits, some level or redundancy appears to be unavoidable (e.g., Reise & Waller, 2009). As for the cures, when the causes are irreducible, trying to achieve an unattainable level of simplicity is likely to result in a loss of accuracy, validity, and utility of the scores. A balance between simplicity and realism must therefore be struck, and, in some cases, this involves explicitly modeling the local dependencies.

So far, local item dependencies in FA have been mostly considered at the structural stage, and a variety of approaches that allow solutions with correlated residuals to be appropriately fitted are available at present (e.g., Ferrando et al., 2025). Their use avoids the problems of misspecification and bias that are expected to occur when items are calibrated by wrongly assuming they are fully locally independent (Mulaik, 2010). However, the impact of the local dependencies at the scoring stage once a correct first-stage structural solution has been fitted has been far less studied, and this is a main issue in this proposal.

The same scoring focus is taken with regard to the second topic above. The problem of deciding between a multiple correlated-factors solution or an essentially unidimensional solution is pervasive in real noncognitive applications, because many data sets conform acceptably with both types of solutions (Calderon et al., 2019; Furnham, 1990; see the ''Added-Value Assessment'' section for a further discussion). Again, this problem has been mostly attacked at the structural level, generally relying on measures of model-data fit (see e.g., Raykov & Calvocoressi, 2021). However, while the degree of model-data fit is indeed a relevant source of information, we believe again that the main sources for deciding the most appropriate level of scoring (multiple subscale scores or total scores) should: (a) be, to a greater or lesser extent, related to the properties of the scores derived from the structural solution; (b) depend as little as possible on the specific method of structural estimation (Raykov & Calvocoressi, 2021); and (c) consider the practical consequences of using one or the other of the two levels in real applications (Reise & Haviland, in press). In this article, we will focus on three internal sources of evidence (discussed in detail in the ''Total Test Analyses'' and ''Added-Value Assessment'' sections), which agree with these conditions: factorial saturation/fidelity, marginal accuracy, and added value.

We shall finally discuss the issue of sum scores, which has been extensively debated in recent years (e.g., Raykov et al., 2015; Sijtsma et al., 2024; Widaman & Revelle, 2023). If the psychometric analysis is based on a two-stage FA application, then the most theoretically defensible type of scores are factor score estimates (or predictors) that use all the information available from the structural stage (e.g., Beauducel & Leue, 2013; Comrey & Lee, 1992). Compared to them, sum scores are suboptimal (although robust) proxies, and their use is expected to result in a loss of accuracy and information (Raykov et al., 2015).

In practice, however, things are more complex. To start with, for the factor score estimates to manifest their theoretical advantages, the structural solution on which

they are based must be really strong, stable, and replicable. Otherwise, these scores are likely to reflect (possibly to a large extent) the sampling error of the item structural estimates (e.g., Sijtsma et al., 2024; Wainer, 1976). Apart from that, the simple sum scores have useful properties. First, they are easy to compute, interpret, and relate to previous results, which make them the most widely used scoring choice in psychometric applications (e.g., Raykov & Marcoulides, 2011; Widaman & Revelle, 2023). Second, they are quite robust as estimates of the factor levels, and, as expected, might provide more stable results under cross-validation (Sijtsma et al., 2024; Wainer, 1976). Finally, and highly relevant for the present proposal, its use does not add additional biases in the estimation of the ''true'' factor levels when there are correlated residuals (this point is demonstrated below). Overall, we do not claim that sum scores are the best choice in the present scenario, but only that they have enough useful properties to justify the interest of the present proposal.

Aims and Contributions

The FA-based general approach we aim to propose is intended to be used in noncognitive psychometric applications in which: (a) non-negligible local item dependencies exist and have been modeled explicitly in the structural solution, (b) the item scores conform to a correlated-factors solution but also show evidence of essential unidimensionality, and (c) the chosen type of scores are sum scores derived from the structural solution. As further discussed below, this is a plausible scenario in personality, attitude, and clinical measurement (Calderon et al., 2019; Furnham, 1990).

The proposal has three general aims. First is to provide tools for assessing the properties of factorial saturation/fidelity, accuracy, and effectiveness of the sum scores obtained at the subscale level and derived from the correlated-factor solution. This first aim is a direct extension of Ferrando's et al. (2025) proposal, and it is based on two general types of measures: reliability and information.

The second aim is to extend the scale-by-scale procedure above to the assessment of the properties of the total test scores as measures of the general factor that is thought to be common to all of the test items. The assessed properties and measures are the same as above. However, the developments are now based on a second-order FA solution and are, to the best of our knowledge, new.

Finally, the third aim is to adapt the added-value principle (Haberman & Sinharay, 2010) to the present scenario and so provide an objective source of information for helping to decide which level of scoring (total or subscale) is the most appropriate for the application at hand. While the added-value principle is indeed well known, its adaptation to the present scenario leads to new results.

Apart from the main aims above, the proposal has also instrumental and illustrative aims. Thus, at the instrumental level, everything that is proposed here has been implemented in a noncommercial, freely available program that is described below. At the illustrative level, finally, an empirical example is used to illustrate the distortions that doublets may cause, and how the various measures considered in this proposal should be interpreted.

Overview: Framework, Requirements, and Intended Use

The structural FA solution we shall consider is a first-order correlated factor compatible with a second-order solution with a single general factor. The solution is assumed to be strong and replicable and have an acceptable degree of model-data fit. So, it can be meaningfully taken as a basis for obtaining the two types of sum scores mentioned in the first, introduction, section.

The ''ideal'' first-order solution for the items to be univocally assigned to subscales is an independent-cluster (IC) solution (e.g., McDonald, 2000). This solution, however, might be unrealistic in many practical applications. So, the ''minimal'' type of primary solution that we shall require here is an independent-cluster-basis (ICB) solution in which at least two items for the primary factor are factorially pure (McDonald, 2000). Furthermore, for the complex items, we shall assume that they have a clear dominant loading on one of the primary factors.

We shall also assume that the first-order loadings of the items chosen to form the subscales; the loadings of the items on the general factor, and the second-order loadings of the primary factors on the general factor are all positive. In an IC solution, these conditions can always be attained by appropriate item reversing (McDonald, 2000). In an ICB solution, attaining these conditions is not warranted but seems a realistic goal given the ICB conditions assumed above (e.g., Ackerman, 1996).

The scenario described so far is standard, achievable if a good design is used, and has been considered in many previous proposals (e.g., Ackerman, 1996; McDonald, 2000; Reise & Haviland, in press; Zinbarg et al., 2007). The main novelty here, however, is the inclusion of item LDs in the basis solution.

The second-order structural solution considered here requires at least three primary factors to be identified. With three primary factors, the second-order part is just-identified and the overall fit of this part of the solution cannot be tested. With only two primary factors, the second-order solution is underidentified. However, Zinbarg et al. (2007) have provided workable solutions for this case that can be used in our proposal.

So as to illustrate the basic framework so far discussed, Figure 1 shows a diagrammatic representation of a very simple second-order IC solution with three primary factors and two pairs of correlated specificities (for clarity, the intercepts are omitted).

The general proposal so far summarized is intended to be used with two types of FA solutions. From a more general SEM view, the first type corresponds to the ''traditional'' SEM, in which the observed variables are treated as continuous-unbounded and all the relations are assumed to be linear (e.g., Raykov & Marcoulides, 2011, 2015). The second type corresponds to the subclass of SEMs based on the underlying-variables-approach (UVA; e.g., B. Muthén, 1984, Raykov & Marcoulides, 2015), in which the item scores are treated as ordered-categorical variables by assuming the existence of an underlying strength latent variable that is discretized at given thresholds (see below for further discussion). The subclass of SEMs based on the UVA is usually estimated using limited-information procedures, generally based on the least-squares principle (e.g., B. Muthén, 1984, 1993). Overall, the framework provided in this section is expected to

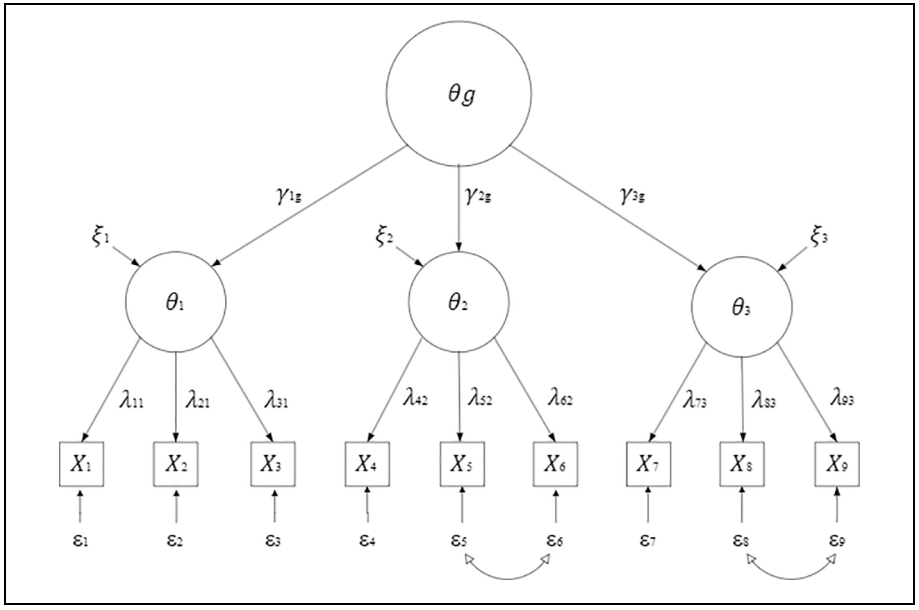


Figure 1. A Simple Second-Order Solution With Correlated Residuals.

hold directly for the observed item scores when fitting ''traditional'' linear solutions, whereas it is expected to hold for the ''strength'' response variables that underlie the ordered-categorical observed item scores when fitting UVA-based solutions.

As mentioned above in the first paragraphs of the manuscript (before the ''Aims and Contributions'' section), structural solutions of the type considered in this section can be fitted by using standard SEM procedures as implemented in widely known available programs (e.g., Raykov & Marcoulides, 2015) and this feasibility holds for both linear ''traditional'' solutions and UVA-based solutions. As a starting point, we shall then consider that the practitioner or the researcher has appropriately fitted a basis structural solution of this type to his or her data. From here, the focus on the proposal that follows is on the appropriate assessment of the properties of the sum scores derived from this solution. As mentioned above, these properties are usually assessed by (wrongly) assuming the items to be fully locally independent even when the local dependencies have been explicitly modeled at the structural level.

Procedures Based on Linear FA Solutions

The General Structural Model

Consider a multidimensional instrument made up of $j=1. . . n$ items, intended to measure a set of $k = 1 . . . m$ common primary factors each of them denoted as θ_k. For a randomly selected respondent i, the basic first-order FA equation in the population is:

$$X_{ij} = \mu_j + \lambda_{j1}\theta_{i1} + \ldots \lambda_{jm}\theta_{im} + \psi_j\, \varepsilon_{ij}, \quad (1)$$

where X_{ij} is a single value of an observed variable: the item score of respondent i on item $j.$; μ_j is an intercept term and is a structural parameter, λ_{jk} is also a structural parameter and is the item loading on the k primary factor, θ_{ik} is a single value of a random latent variable: the level, or factor score of the respondent i in the k primary factor; ψ_j is the item residual standard deviation and is also a structural parameter; and, finally, ε_{ij} is a single value of a latent random variable (e.g., Raykov & Marcoulides, 2011, p. 40): the residual score of respondent i on item j. The term residual is justified because it can be viewed as the residual score of item j about its regression on the common factors (McDonald, 1981, p. 106). At the conceptual level, this residual can be further decomposed into two parts: specificity (i.e., consistent sources of variance not captured by the common factors) and random measurement error (e.g., McDonald, 1981; Raykov & Marcoulides, 2011). The correlated residuals we consider here arise because of relations between the specific parts of the residuals.

With regard to scaling and assumed relations, the common factors and the residual scores are in standard scale (zero mean and unit variance). So, the structural parameters, λ_{jk}. and ψ_j can be also viewed as ''scaling'' parameters that bring the standardized common factor and residual scores to the same scale as that of the observed item scores. Finally, the residual scores are assumed to be uncorrelated with the primary factors and some of the residuals are correlated among them (see Figure 1).

For each primary factor, the second-order FA equation is next given by:

$$\theta_{ik} = \gamma_{kg}\theta_{ig} + \xi_{ik}, \quad (2)$$

where γ_{kg} is a structural parameter: the second-order loading of the k primary factor on the second-order general factor θ_g; θ_{ig} is a single value of a random latent variable: the level of the respondent i in the general factor; and ξ_{ik} is a single value of a latent random variable: the residual associated with the latent primary factor as an indicator of the general factor and denoted as structural residual (e.g., Rindskopf & Rose, 1988). The structural residuals ξ_{ik} are assumed to be uncorrelated both with the general factor and among them. The general factor is also scaled on a standard scale.

Using vector–matrix notation, the first-order Equation 1 can be written as:

$$\mathbf{x} = \boldsymbol{\mu} + \boldsymbol{\Lambda}\ \boldsymbol{\theta} + \boldsymbol{\Psi}\boldsymbol{\varepsilon}, \quad (3)$$

where $\mathbf{x}$ and $\boldsymbol{\varepsilon}$ are the $n\times 1$ random vectors of observed item scores and residual scores, respectively, $\boldsymbol{\theta}$ is an $m\times 1$ random vector of factor scores, $\boldsymbol{\mu}$ is an $n\times 1$ vector of intercepts, $\boldsymbol{\Lambda}$ is an $n\ x\ m$ matrix (pattern) of first-order loadings, and $\boldsymbol{\Psi}$ is an $n\ x\ n$ diagonal matrix containing the item residual standard deviations.

The second-order solution (2) in vector–matrix notation is:

$$\boldsymbol{\theta} = \boldsymbol{\Gamma}\boldsymbol{\theta}_{\boldsymbol{g}}\ + \boldsymbol{\xi} \quad (4)$$

where $\boldsymbol{\Gamma}$ is an $m\times 1$ vector of second-order loadings and ξ is an $m\times 1$ vector of structural residuals.

Finally, the covariance structure implied by Equations 3 and 4 is:

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi} \\ &= \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Delta}_{\xi\xi}\right)\boldsymbol{\Lambda}' + \boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}, \end{aligned} \quad (5)$$

where $\boldsymbol{\Delta}_{\xi\xi}$ is the diagonal m x m structural-residual covariance matrix and $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ is the n x n item residual correlation matrix. The term in parentheses in the second expression in Equation 5 is the correlation matrix between the primary factors (e.g., Rindskopf & Rose, 1988), and, since the structural residuals are assumed to be uncorrelated, these primary factors correlate only due to their weights on the second-order factor (e.g., Rindskopf & Rose, 1988).

If all the items are locally independent, $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ reduces to an identity matrix, and the following structure is obtained:

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}' + \boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}^2 \\ &= \boldsymbol{\Lambda}\left(\boldsymbol{\Gamma}\boldsymbol{\Gamma}' + \boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\right) + \boldsymbol{\Psi}^2, \end{aligned} \quad (6)$$

(see, e.g., Rindskopf & Rose, 1988, Equation 3, p. 53; and please note that there is only a single second-order factor here).

It is noted that Equations 5 and 6 are the covariance structures and, because covariances are obtained by centering the variables, the intercept terms in Equations 1 and 3 play no further role at the structural level.

The covariance structure (6) is generally falsifiable, and the degree of model-data fit can be assessed using available SEM procedures. When item residual correlations are allowed to be freely estimated (i.e., Equation 5), one degree of freedom will be lost per correlation, and so, if the number of allowed residual correlations was large enough, there may not be enough degrees of freedom available to estimate and test the solution. In the present proposal, we will not consider this scenario. Rather, we will assume that only a certain number of correlated residuals are specified as free and so that the proposed solution is identifiable and testable.

As a summary, the modeling described in this section is based on the common assumptions which are made in ‘‘traditional’’ SEM, mainly, that all the regressions are linear (which, in this case, implies the regressions of the item scores on the primary factors and those of the primary factors on the general factor) and that the residual variances (both measurement and structural) are homoscedastic (e.g., McDonald, 1985, 1999).

With regard to the assumptions concerning the item correlated residuals, we focus on the bivariate level and conceptualize local independence according to the ‘‘weak form’’ of the LI principle (McDonald, 1985, 1999): that all the correlations among residuals vanish once the influence of the common factor/s has been partialled-out. The ‘‘strong form’’ of LI would further require all the higher-order joint moments

(not only the second-order correlations) would also vanish. In practice, however, McDonald clearly states that the weak form adopted here largely suffices in applications. Finally, we acknowledge that LDs could be modeled in alternative ways to that proposed here even within the FA model, mainly by treating them as method effects (e.g., Schweizer et al., 2024).

Separate Subscale Analyses

This level of analysis is equivalent to the single-factor assessment proposed by Ferrando et al. (2025). For each of the *1..k..m* primary factors, consider the subscale score S_k obtained by summing the scores of the n_k items that are defined to be the indicators of θ_k. Let $\mathbf{1_k}$ be an $n_k \times 1$ unit vector using obvious notation for the remaining terms, the conditional mean and variance of S_k for fixed θ_k are:

$$E(S_k|\theta_k) = 1'_k\mu_k + 1'_k\boldsymbol{\lambda_k}\theta_k.$$

$$Var(S_k|\theta_k) = 1_k'\boldsymbol{\psi}'_k\boldsymbol{\Sigma}_{\varepsilon\varepsilon(\mathbf{k})}\boldsymbol{\psi_k}\,1_k. \tag{7}$$

So, whether there are correlated residuals or not, the subscale sum score is an unbiased estimate of a fixed linear function of θ_k and its conditional variance does not depend on θ_k. The corresponding marginal mean and covariance are:

$$E(S_k) = 1'_k\boldsymbol{\mu_k}.$$

$$Var(S_k) = 1'_k\boldsymbol{\lambda_k}\boldsymbol{\lambda}'_k1_k + 1'_k\boldsymbol{\psi}'_k\boldsymbol{\Sigma}_{\varepsilon\varepsilon(\mathbf{k})}\boldsymbol{\psi_k}\,1_k. \tag{8}$$

Using Lord's (1980) definition, the information contributed by S_k as implied by Equation 7 is:

$$I_{LD}(\theta_k, S_k) = \frac{\left(\frac{dE(S_k|\theta_k)}{d\theta_k}\right)^2}{Var(S_k|\theta_k)} = \frac{1'_k\boldsymbol{\lambda_k}\boldsymbol{\lambda}'_k1_k}{1_k'\boldsymbol{\psi}'_k\boldsymbol{\Sigma}_{\varepsilon\varepsilon(k)}\boldsymbol{\psi_k}\,1_k}. \tag{9}$$

The squared correlation between S_k and θ_k is:

$$\rho^2(\theta_k, S_k) = \frac{1'_k\boldsymbol{\lambda_k}\boldsymbol{\lambda}'_k1_k}{1'_k\boldsymbol{\lambda_k}\boldsymbol{\lambda}'_k1_k + 1'_k\boldsymbol{\psi}'_k\boldsymbol{\Sigma}_{\varepsilon\varepsilon(k)}\boldsymbol{\psi_k}\,1_k} = \omega_{LD(k)}. \tag{10}$$

And is an expression of the omega reliability coefficient (McDonald, 1985) when correlated residuals exist (e.g., Raykov, 2001).

The relation between Equations 9 and 10 is then

$$\omega_{LD(k)} = \frac{1}{1 + \frac{1}{I_{LD}(\theta_k, S_k)}};\; I_{LD}(\theta_k, S_k) = \frac{\omega_{LD(k)}}{1 - \omega_{LD(k)}}. \tag{11}$$

And, in this case, it is one-to-one. Conceptually, however, both indices are intended to address different properties of the scores. The information measure is mainly a signal/noise ratio (see Cronbach & Gleser, 1964) which indicates how many times the variance of the subscale scores associated with the primary factor is larger than the error variance of these scores. For its part, $\omega_{LD(k)}$ can be viewed both as a reliability coefficient (if the θ_k levels are considered to be ''true'' scores) and as an index of factor saturation (e.g., Reise et al., 2010; Revelle & Zinbarg, 2009) that reflects (in this case) the percent of variance in the subscale sum scores that can be explained by the primary factor that they intend to measure. Finally, the square root of $\omega_{LD(k)}$ is a fidelity coefficient (Drasgow & Miller, 1982): the product–moment correlation between the subscale scores and the factor they intend to measure. So, it indicates the degree of association these scores have with their primary factor. Apart from these conceptual distinctions, however, and even more important for the developments that now follow, is the result that the information measure (9) has no upper bound, whereas the coefficient (10) has a unit upper bound.

Consider now the situation in which the subscale items would maintain the same structural properties but were fully locally independent. The information and reliability measures would now be obtained by Equations 9 and 10 but by using and identity matrix instead of $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon(k)}}$. These results, denoted by $I_{LI}(\theta_k, S_k)$ and $\omega_{LI(k)}$, respectively, predict then the amounts of information and accuracy that could be attained if the items that are calibrated were locally independent.

Consider finally the ratio:

$$RE_{LD(k)} = \frac{I_{LD(\theta_k, S_k)}}{I_{LI(\theta_k, S_k)}} = \frac{1_{\boldsymbol{k}}{}'\boldsymbol{\psi}'_{\boldsymbol{k}}\,\boldsymbol{\psi}_{\boldsymbol{k}}\,1_{\boldsymbol{k}}}{1_{\boldsymbol{k}}{}'\boldsymbol{\psi}'_{\boldsymbol{k}}\,\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon\varepsilon(k)}}\boldsymbol{\psi}_{\boldsymbol{k}}\,1_{\boldsymbol{k}}} = \frac{\omega_{LD(\theta_k, S_k)}\left(1 - \omega_{LI(\theta_k, S_k)}\right)}{\omega_{LI(\theta_k, S_k)}\left(1 - \omega_{LD(\theta_k, S_k)}\right)}. \tag{12}$$

Expression 12 is a relative efficiency measure (e.g., Lord, 1980) and quantifies the change (generally loss) of information that is due to the local dependencies among items. Furthermore, the expression at the right end of Equation 12 shows that *RE* can also be interpreted in terms of test length. So, if *RE* is lower than 1 (as generally expected), the result can be interpreted as that the scores on the subscale with LDs measure the primary factor with an accuracy equivalent to that which the locally independent scale would have if the latter had been shortened *RE* times. To give a simple example, if the subscale has 100 items and the *RE* in Equation 12 is 0.8, then the loss in accuracy/information due to the local dependencies would be equivalent to a loss of 20 items.

Total Test Analyses

Let now S be the total score obtained by summing all the $j=1. \ldots n$ item scores, and taken to be a measure of the general second-order factor θ_g. By following the same developments as in the previous section, we obtain now

$$E\left(S|\theta_g\right) = 1'\boldsymbol{\mu} + 1'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\theta_g.$$

$$Var(S|\theta_g) = \mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}\ \mathbf{1}. \quad (13)$$

And

$$E(S) = \mathbf{1}'\boldsymbol{\mu}.$$

$$Var(S) = \mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}\ \mathbf{1}. \quad (14)$$

So, again, the total sum scores are unbiased estimates of a fixed linear function of θ_g and their conditional variance does not depend on θ_g.

The information contributed by the total score S as a measure of the general second-order factor θ_g is:

$$I_{LD}(\theta_g, S) = \frac{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\mathbf{1}}{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}\ \mathbf{1}}. \quad (15)$$

And the squared correlation between S and θ_g is:

$$\rho^2(\theta_g, S) = \frac{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\mathbf{1}}{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Gamma}\boldsymbol{\Gamma}'\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}\ \mathbf{1}} = \omega_{LD(H)}. \quad (16)$$

Expression 16 is an extension of the omega hierarchical coefficient (McDonald, 1999; Revelle & Zinbarg, 2009) when derived from a solution with correlated specificities. Conceptually, this coefficient estimates the percent of variance of the total sum scores associated with only the general second-order factor (see Reise & Haviland, in press for a discussion). It can also be considered to be a reliability coefficient if the general second-order factor levels are viewed as true scores. Finally, its square root can also be interpreted as a fidelity coefficient.

The relation between Equations 15 and 16 has the same form as in the scale-by-scale case:

$$\omega_{LD(H)} = \frac{1}{1 + \frac{1}{I_{LD}(\theta_g, S)}};\ I_{LD}(\theta_g, S) = \frac{\omega_{LD(H)}}{1 - \omega_{LD(H)}}. \quad (17)$$

The limiting predictions if the total item set was fully locally independent can be readily obtained again by using and identity matrix instead of $\boldsymbol{\Sigma}_{\varepsilon\varepsilon}$ in Equations 15 and 16 and will be denoted by $I_{LI}(\theta_g, S)$ and $\omega_{LI(H)}$, respectively. The relative efficiency is:

$$RE_{LD} = \frac{I_{LD(\theta_g, S)}}{I_{LI(\theta_g, S)}} = \frac{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Psi}\ \mathbf{1}}{\mathbf{1}'\boldsymbol{\Lambda}\boldsymbol{\Delta}_{\xi\xi}\boldsymbol{\Lambda}'\mathbf{1} + \mathbf{1}'\boldsymbol{\Psi}\boldsymbol{\Sigma}_{\varepsilon\varepsilon}\boldsymbol{\Psi}\ \mathbf{1}} = \frac{\omega_{LD(H)}(1 - \omega_{LI(H)})}{\omega_{LI(H)}(1 - \omega_{LD(H)})}, \quad (18)$$

and has an interpretation analogous to that in the previous section. In this case, the relative loss of information/accuracy of the total scores as measures of the general

factor is due to the LDs. The corresponding interpretation in terms of test length also holds.

In closing this section, some comments are in order. Cut-offs or reference values for interpreting omega hierarchical coefficients and fidelity coefficients have been previously proposed, but they were intended for fully locally independent solutions. For example, Revelle (1979) suggested a .50 cutoff in omega-H to justify the calculation of the total score, and Drasgow and Miller (1982) considered that, ideally, the fidelity coefficients should be above .90 for the sum scores to provide univocal measurement of the intended factor. In light of the present developments, these values should be considered for $\omega_{LD(H)}$ (and its square root) not for $\omega_{LI(H)}$ that is routinely used by default in applications.

Added-Value Assessment

As mentioned above, the present proposal is intended for a scenario in which a correlated-factor solution fits well and is interpretable, but, at the same time, there is also clear evidence of a general factor running through all of the items. At the structural level, the typical results corresponding to this scenario are the following. First, the multiple solution fits better than the unidimensional solution in pure goodness-of-fit terms (e.g., Furnham, 1990). Second, however, the indicators mentioned above that are not specifically related to the estimation method suggest that there is essential unidimensionality. Thus, for example, the percentage of common variance explained by the general factor is high and its increase when going from one to two factors is small, the unidimensional structure has a positive manifold, with all the items loading positively and substantially on the first canonical factor, and the strength and potential replicability of the solution as measured by the *H* index is high (see e.g., Ferrando et al., 2024 and Raykov & Calvocoressi, 2021). Faced with this evidence, the researcher must decide which is the most appropriate solution (Furnham, 1990).

As further discussed below, a decision of this type has to be based on multiple sources of evidence, of which Section 1, which is based on the properties of the scores, is proposed. More specifically, we aim to propose here an objective tool for helping the researcher to decide which type of scoring: total or subscale is the most appropriate when local item dependencies exist. It is based on the added-value principle (e.g., Haberman & Sinharay, 2010).

Consider again the subscale scores S_k intended to measure the primary factor θ_k. In the present scenario, the scores S_k would be considered to have added value if they are able to predict θ_k (i.e., its own factor) more accurately than the total scores S (which are intended to measure the general factor θ_g). This result can be operationalized as (see Ferrando & Lorenzo-Seva, 2019):

$$\rho^2(\theta_k, S_k) > \rho^2(\theta_k, S). \tag{19}$$

Or, in the present case:

$$\omega_{LD(k)} > \rho^2(\theta_k, S). \tag{20}$$

By using standard covariance algebra, the model-based squared correlation on the right-hand side of Equations 19 and 20 is found to be:

$$\rho^2(\theta_k, S) = \frac{1'\mathbf{\Lambda t}^{(k)}\mathbf{t}^{(k)'}\mathbf{\Lambda}'\mathbf{1}}{1'\mathbf{\Lambda\Gamma\Gamma}'\mathbf{\Lambda}'\mathbf{1} + 1'\mathbf{\Lambda\Delta}_{\xi\xi}\mathbf{\Lambda}'\mathbf{1} + 1'\mathbf{\Psi\Sigma}_{\varepsilon\varepsilon}\mathbf{\Psi 1}} \tag{21}$$

where $t^{(k)}$ is an *m x 1* vector whose *k*th element is *1* and the remaining elements are the products: $\gamma_{1g}\gamma_{kg}$ . . .. . .: $\gamma_{mg}\gamma_{kg}$.

Provided that the subscale scores significantly demonstrate added value (see below), the difference between the two terms in Equation 19:

$$ES_k = \rho^2(\theta_k, S_k) - \rho^2(\theta_k, S), \tag{22}$$

can be interpreted as an incremental effect-size measure of the explained-proportion-of-variance type (e.g., Cohen, 1992), a measure which is conceptually similar to those proposed by Raykov and Calvocoressi (2021) and Ferrando et al. (2024). In effect, the Difference 22 would be interpreted as the increment in the explained variance of θ_k that is obtained when predicting θ_k from its specific subscale scores instead of from the general scale. Raykov and Calvocoressi (2021) provided initial, tentative guidelines, which agree with those proposed by Cohen (1992), and that we shall also consider here. So, with due reservation, values of ES_k in the range .05 to .08 would be considered small; in the .15 to .25 range medium, and above .30 large.

The proposal in this section is partly based on existing procedures, but it has been explicitly adapted to the situation in which LDs exist. As Equations 19 to 21 show, the outcomes of the proposed indicators could differ from those that would be obtained if full local independence was assumed.

Taking Sampling Error into Account: Confidence Intervals

In their unidimensional proposal, Ferrando et al. (2025) proposed a simulation/resampling procedure that allowed Confidence Intervals (CIs) for all the indices they proposed to be obtained. This is also the approach we shall also consider here. While it is very simple, if the first-stage structural estimation has been carried out according to the requirements discussed above, it is expected to work well.

The basic idea is that proposed by Bollen and Stine (1992): to sample from a population in which the structural model is taken to be correct. In our case, the user will be asked to provide the needed structural point estimates (see Equation 5), together with the size of the sample in which they have been obtained. Next, this solution is taken to be fixed and correct, and used to generate pseudo-samples or replicas of the same size as that specified by the user. Finally, for the indices at the subscale and scale level, the upper and lower limits of the 90% CIs are taken to be the 5th and the 95th percentile of the distribution across replicas.

With regard to the added-value assessment, our proposal is as follows. In each replica, the incremental effect-size measure (22) is computed for each of the primary factors, and the CIs are computed as above. If both limits of the CIs lie above zero, we shall consider that the subscale scores have added value, and the point estimated *ES* in Equation 22 will be interpreted according to the tentative guidelines above.

It is important to note, finally, that in the approach so far discussed, the point estimates of the indices are first obtained analytically, by using directly the structural estimates (see Equations 10 to 22). However, the CIs are obtained via simulation, using the structural estimates as a basis for data generation. In the test studies, we carried out the average values of the indices across replicas were extremely close to the corresponding analytically derived point estimates. These results suggest that point estimates and CIs could be all jointly obtained by using the approach described in this section.

The Extension to the Nonlinear UVA-FA Case

Ferrando et al. (2025) derived indices of score accuracy and effectiveness for unidimensional UVA-FA solutions on the basis of two conditions. First, all the indices were derived from the squared-correlation definition in Equation 10. Second, they were all based on the first-stage structural estimates (i.e., model based). The same conditions can be used for the extended procedures here.

We shall start by reviewing the features of the UVA-FA approach (e.g., B. Muthén, 1984) that are most relevant for the developments that follow. For each item, the UVA model assumes that there is an underlying, continuous-unbounded ''strength'' latent variable that generates the observed item categorical score. These strength UVs are further assumed to be normally distributed with zero mean and unit variance and the distribution of the residuals εs is also assumed to be multivariate normal with correlation matrix $\mathbf{\Sigma}_{\boldsymbol{\varepsilon\varepsilon}}$. Denote by Y_{ij} the latent score of individual i in the latent variable that underlies item j and by X_{ij} the corresponding observed score, which, in a response format with c categories, is obtained by assigning integer values 1, 2 . . . c to the ordered-categorical responses. The process that produces X_{ij} from the Y_{ij} is a nonlinear step function governed by c-*1* thresholds (τ; which are considered to be structural parameters in the model):

$$\begin{aligned} &X=1 \quad if \quad Y<\tau_1 \\ &X=2 \quad if \quad \tau_1 \leq Y<\tau_2 \\ &X=3 \quad if \quad \tau_2 \leq Y<\tau_3 \,. \\ &\ldots \\ &X=c \quad if \quad \tau_{c-1}<Y \end{aligned} \tag{23}$$

The key points regarding the present proposal are two. First, the model defined in Equations 1 to 6 is assumed to hold here for the strength UVs *Y's* (and this includes the basic assumptions discussed above). Second, the sum scores (both subscale and

total) are obtained by adding the observed item scores *Xs*. With regard to the first point, it is worth noting that, because the UVs are standard variables: (a) the intercepts in Equations 1 and 3 are all zero, (b) the first-order loadings in Equations 1 and 3 are standardized loadings, and (c) the covariance matrix $\boldsymbol{\Sigma}$ in Equation 5 becomes now a polychoric correlation matrix.

Denote by S_{Xk} the subscale score obtained by summing the observed scores of the n_k items that are defined to be the indicators of θ_k, and by S_X the total score, obtained by summing all the observed item scores, and considered an indicator of the general factor θ_g. The basis indices in the UVA extension are:

$$\rho^2(\theta_k, S_{Xk},) = \omega_{LD(k)-UVA}. \tag{24}$$

At the subscale level, and

$$\rho^2(\theta_g, S_X) = \omega_{LD(H)-UVA}. \tag{25}$$

At the total test level.

The Indices 24 and 25 have the same reliability-saturation-fidelity interpretation as in the linear case, and, for this reason, they have all also been labeled as omega coefficients. When interpreting them, however, two important points should be taken into account. First, they are based on structural parameters derived from a model that holds for the UVs, whereas the subscale and test scores are obtained from the sums of the observed item scores. So, for a given set of structural parameters, Equations 24 and 25 will be attenuated with respect to those that would be obtained from the continuous variables. As Equation 23 shows, this attenuation would reflect the impact of both categorization (coarse grouping) and nonlinearity. Second, unlike what occurs in the linear model, in the UVA model, both the reliability and the information vary generally as a function of the factor levels. So, if the coefficients in Equations 24 and 25 are to be interpreted as reliability coefficients, they should be viewed as estimates of the marginal reliabilities that would be obtained by averaging the conditional score reliabilities across all the corresponding factor levels.

Once Equations 24 and 25 have been estimated, the information measures can be easily obtained by using the generic relations (e.g., Nicewander, 1993):

$$\omega_{LD-UVA} = \frac{1}{1 + \frac{1}{I_{LD-UVA}}};\ I_{LD-UVA} = \frac{\omega_{LD-UVA}}{1 - \omega_{LD-UVA}}. \tag{26}$$

And it can be interpreted as the inverse of the averaged squared standard errors of measurement when the sum scores (subscale or total) are used for estimating the corresponding *θs*. From here, the remaining measures proposed in the linear case (limiting estimates under local independence and relative efficiency) follow directly.

Analytical expressions for Equations 24 and 25 can be obtained by using the relations between the polyserial and the point-polyserial correlations (see Ferrando et al., 2025, Equations 16 to 19). The resulting results, however, are computationally

complex. So, it seems more practical to jointly obtain the point and CI estimates of Equations 24 and 25, as well of those of the derived indices, by using the simulation approach described in the previous section. This approach also provides estimates of the added value and *ES* measures by estimating directly the squared correlations (19) but with the sum scores as defined here.

Implementation: The Program SINRELADD.LD

The proposal so far discussed has been fully implemented as R script called SINRELADD.LD (Score information, reliability, relative efficiency, and added value in multidimensional measures that contain locally dependent items.). SINRELADD.LD has been developed in R Version 4.4.1 and runs with R versions more recent than 3.5.0. The program uses as input the calibration item estimates obtained from fitting extended second-order FA-solutions, in which the existing LDs are included.

The program is released in a compressed folder, which includes two main R scripts: SINRELADD.LD and SINRELADD.LD. GUI. In both cases, the user must source the desired function before its utilization.

The first script, SINRELADD.LD, was designed for advanced R users, where the input values have to be provided through code and the output is also printed in the R console. A description of the input values is provided in the code itself, and the usage is the following:

```
> SINRELADD.LD(LAM, GAM, DAT, model = ''linear,'' method = ''point-estimate,''
doublet_list, cor_doublet, N)
```

The second script, SINRELADD.LD.GUI is a shiny app (Chang et al., 2024) which provides a Graphical User Interface version of the script, more suitable for users less familiar with R language. It requires the shiny package to be installed, and the usage is the following:

```
> library(shiny); runApp(''SINRELADD.LD.GUI.R'')
```

The uploaded version contains a detailed user's guide in addition to the already existing documentation embedded in the R package. Finally, a data folder is also provided, containing an example data set and all the required input data for using SINRELADD.LD.

The program can be downloaded from: https://psico.fcep.urv.cat/utilitats/SINRELADD-LD/.

Empirical Example

We have used the data of the 928 respondents (85.7% females) who participated in the study by Dueñas et al. (2022). Their ages ranged between 28 and 69 years old (M = 46.86, SD = 5.45). They answered the Spanish adaption of the Family Involvement Questionnaire-High School Version (FIQ-HS), which evaluates the parental involvement in their teenage children's education. The FIQ-HS is a 33-item questionnaire with a 4-point Likert-type response format (1 = *Rarely*, 4 = *Always*) aimed at assessing three factors: Home-school communication (10 items), school-based activities (6 items), and home-based activities (17). The first factor (F1) refers to the forms of contact that parents might have with school staff (e.g. talking with teachers about difficulties at school). The second factor (F2) refers to parent behavior in the school setting (e.g. volunteering, participating in school fundraising activities, etc.). The third factor (F3), finally, refers to parental activities outside school that promote learning (e.g. helping teenage children with homework). Dueñas et al. (2022) found that the error terms of three pairs of items belonging to the third factor were substantially correlated. According to these authors, the item stems of each pair have a very similar wording or content, which would explain this result.

In the present study, a second-order UVA-FA solution with the three primary factors explained above and a general factor of parent involvement was specified. Furthermore, in this specification, the three correlated residuals above were freely estimated. The solution was fitted to these data by using robust ULS estimation as implemented in Mplus 8.10 (L. K. Muthén & Muthén, 2017). Goodness-of-fit results were quite acceptable: Root Mean Square Error of Approximation (RMSEA) = 0.037, 90% CI = [0.034, 0.040] Comparative Fit Index (CFI) = 0.92; Goodness of Fit index (GFI) = 0.97. Note, however, that with only three primary factors, the second-order part of the solution is just-identified and the fit is the same as that obtained from the correlated-factors solution. The standardized item loading estimates are in Table 1, and, according to them, the first-order structure is a clean, fairly strong IC solution. Regarding the second-order pattern, Table 1 shows the gamma estimates of each primary factor on the second-order factor. As can be seen, the three primary factors are all good indicators of the general factor, but F1 is the strongest and more g-saturated indicator.

The three doublets mentioned above, which are located in F3, were found to be quite strong. More specifically, the estimated residuals correlations were: .82 for the pair 8 to 21, .78 for the pair 25 to 26, and .49 for the pair 27 to 31.

We shall now focus on the results at the separate subscale level. For each subscale, Table 2 shows both the correct Omega reliability estimate (ω_{LD}), in which the local dependencies are considered, and the ''ceiling'' estimate if the items were locally independent (ω_{LI}). As it should be, differences were only found in the third subscale. Although both Omega values seem to be very similar, their CIs do not overlap, which involves that ω_{LD} is significantly smaller than ω_{LI}. Furthermore, ω_{LD} does not include the .80 value commonly considered as a minimum threshold (e.g., Raykov & Marcoulides, 2011). The Score Relative efficiency of .83 suggests that the

Table 1. Standardized Item Loadings Estimates for the First-Order Pattern, and Gamma Estimates of Each Factor for the Second-Order Pattern.

First-order pattern			
Item	F1	F2	F3
1	**.60**	.00	.00
2	**.70**	.00	.00
7	**.69**	.00	.00
11	**.72**	.00	.00
12	**.76**	.00	.00
16	**.78**	.00	.00
22	**.62**	.00	.00
23	**.59**	.00	.00
30	**.62**	.00	.00
32	**.54**	.00	.00
5	.00	**.64**	.00
6	.00	**.70**	.00
14	.00	**.81**	.00
15	.00	**.77**	.00
17	.00	**.76**	.00
29	.00	**.79**	.00
3	.00	.00	**.39**
4	.00	.00	**.48**
8	.00	.00	**.68**
9	.00	.00	**.60**
10	.00	.00	**.65**
13	.00	.00	**.69**
18	.00	.00	**.59**
19	.00	.00	**.71**
20	.00	.00	**.51**
21	.00	.00	**.75**
24	.00	.00	**.62**
25	.00	.00	**.38**
26	.00	.00	**.39**
27	.00	.00	**.58**
28	.00	.00	**.72**
31	.00	.00	**.56**
33	.00	.00	**.65**
Second-order pattern			
	Gamma 1	Gamma 2	Gamma 3
	0.93	0.46	0.57

Note. F1 = Home-school communication; F2 = School-based activities; F3 = Home-based activities. Significant values are in bold ($p < .01$).

loss of accuracy due to the doublets is equivalent to that which would occur if 17% of items were removed.

Table 2. Omega Reliability Estimates Under Local Independence and Under Local Dependency for Each Subscale (90% Confidence Intervals), Score Relative efficiency, and Coefficient of Fidelity.

	S1	S2	S3
Omega-LD	.84 [.83, .85]	.76 [.74, .78]	.77 [.76, .78]
Omega-LI	.84 [.83, .85]	.76 [.74, .78]	.80 [.79, .81]
RE	1	1	.83 [.77, .90]
COF-LD	.92 [.91, .93]	.87 [.86, .88]	.88 [.87, .89]

Note. RE = Score Relative efficiency; COF-LD: Coefficient of Fidelity (square root of ω_{LD}); S1 = Home-school communication subscale; S2 = School-based activities subscale; S3 = Home-based activities subscale.

As explained above, the square root of ω_{LD} is a fidelity coefficient that can be interpreted as the estimated correlation between the subscale scores and the factor they intend to measure. Table 2 shows the Coefficient of Fidelity (COF-LD) for each subscale. As can be seen, once LDs have been taken into account, all COF-LD estimates are higher than .85, and one of them (Home-school communication) is above the .90 threshold suggested by Drasgow and Miller (1982) (note also that the entire CI is above .90). Overall, all the subscale scores can be considered to be acceptably good proxies of the corresponding factor levels.

The following step focuses on the second-order results. The $\omega_{H\text{-}LD}$ estimate is .66; 90% CI = [.64, .68], which is clearly above the minimum Revelle's proposed threshold for meaningfully using total scores. The $\omega_{H\text{-}LI}$ is slightly higher: .67; 90% CI = [.64, .69]. The Score Relative efficiency estimate is .98, 90% CI = [.92, 1]. Therefore, the loss of accuracy here does not reach statistical significance and, in any case, it would be very small. This is mainly due to the fact that F1 (which is the best first-order indicator of the general factor) has no correlated specificities. In fact, the three modeled doublets are part of F3, which is a far weaker indicator of the general factor than F1.

The last step focuses on the added-value assessment and effect-size results of the subscale scores, in comparison with the overall scores. Table 3 shows the estimated effect sizes and the 90% CIs. All of them differ significantly from 0, according to their CIs. So, in all cases, there is added value when subscale scores are used instead of overall scores to predict their own factor. Using the tentative guidelines explained above, the effect size for F2 would qualify as large, while the effect sizes for F1 and F3 are medium. In other words, F2 is the factor with the largest incremental effect size when the subscale scores are used instead of the overall scores for predicting it. This result is not surprising, given that F2 is the weakest indicator of the general factor (see the gamma estimates in Table 1). As F1 is a very good indicator of the general factor, whether subscale or overall scores are used does not make as big a difference as in the case of F2. For F3, the increase is also moderate, partly because

Table 3. Effect-Size Results and 90% Confidence Intervals.

	Effect size
F1	.20 [.18, .22]
F2	.43 [.39, .46]
F3	.16 [.13, .18]

Note. F1 = Home-school communication subscale; F2 = School-based activities subscale; F3 = Home-based activities subscale.

of the presence of three doublets and partly because this factor is not an excellent indicator of the general factor. If the items in F3 had been locally independent, the estimated ES would have risen from .16 to .19 (exceeding the CI in Table 3).

Discussion

This article focuses on several relatively common problems in the assessment of noncognitive attributes such as personality, attitudes, or clinical characteristics. The first one is the presence of items with correlated residuals that share specific, noncontent variance. In some cases, this problem cannot be solved by simply eliminating some offending items (e.g., Reise & Waller, 2009) and it becomes necessary to model correlated residuals by using extended FA solutions that include them. However, the implications of this modeling go beyond the structural level, because the presence of LDs affects the accuracy and efficiency of scores derived from the extended solutions.

In the present proposal, we have focused on sum scores, and we first assessed the functioning of subscale scores derived from a primary correlated-factors solution, either linear or UVA. First, we compared the omega reliability estimates derived from taking into account the LDs with those that would be obtained by assuming full local independence. As can be seen in the empirical example, in the subscale with correlated residuals, the accuracy of the scores is overestimated when local independence is assumed, which is consistent with the results in Ferrando et al. (2025). For this reason, when using measures that include correlated residuals, we encourage the use of the omega coefficient that assumes local dependencies, although it may seem counter-intuitive to use a coefficient that will likely result in a lower value than the ordinary omega. However, it is important to have an unbiased idea of the actual accuracy with which scores are estimated; otherwise, wrong decisions may be made when using the scores in individual assessments or on score comparisons (see Wainer & Thissen, 1996 for a detailed discussion about this point). In addition to the reliability assessment, we also proposed to use a Score Relative Efficiency index, which quantifies the magnitude of the loss of accuracy due to LDs and which has a simple interpretation in terms of the percentage of items lost.

A main contribution of this proposal is to first extend the primary solution above to a second-order solution with a single general factor and next assess the functioning of the total scale scores as measures of the general factor that is assumed to run across all the items. The main indicators are the same as those used at the subscale level, but their use here is expected to provide useful additional information. Thus, although the doublets may cause a loss of accuracy in their subscale itself, this effect may be diluted in the total scores, for example, if there are not so many doublets in relation to the total number of items, or if the magnitude of the correlation between their residuals is not relevant enough. Furthermore, as can be seen in the empirical example, their effect can also be diluted if the doublets are concentrated in a factor that is not excessively saturated by the overall factor. In addition to this, the Score Relative Efficiency index for the total scores provides further information about the magnitude of the overall loss of accuracy that goes beyond that obtained at the subscale level.

A second main contribution of the article is to extend the added-value principle to the second-order scenario when LDs are operating and provide an auxiliary tool for judging the usefulness of using subscale scores rather than a single total score in multidimensional measures. This contribution is embedded in the more general issue of deciding whether the correlated-residual solution is, in fact, more appropriate and useful than the essentially unidimensional solution. We believe that a decision of this type requires multiple sources of evidence to be obtained, mainly: (a) internal, structural evidence, (b) properties of the scores derived from the structural solutions, and (c) external validity evidence. The added-value proposal focuses only on the second source and so it is considered only an auxiliary source. However, our measures and, particularly, the effect-size index and its CIs provide useful information. On one hand, if using subscale scores rather than overall scores does not involve any gain, then the usefulness of the subscales can be questioned. On the other hand, if they do, the use of the proposed measures makes it possible to differentiate which subscales show the greatest gains with respect to the total scores.

To sum up, our proposal includes several indices and procedures (some of them new), intended for multidimensional measures with local item dependencies, which provide information about the loss of accuracy in the total scores due to the doublets, and about the magnitude of this loss. The assessment goes beyond that previously proposed at the unidimensional level and provides a more comprehensive view about the consequences of the doublets in the whole instrument. Furthermore, the added-value and effect-size measures are also a relevant contribution, as they may be helpful to decide whether it is worthwhile to calculate the subscale scores separately, instead of using overall scores. Overall, it seems clear that the impact of LDs goes far beyond the item calibration stage and is expected to also affect the properties of the scores. So, developments such as those in this proposal are needed if sound score decisions in measures that include doublets have to be made.

As any relatively new proposal, this one has its share of limitations. Thus, much more evidence is needed for deciding which are the best criteria and reference values

for using the proposed indices. This is particularly relevant in the case of the added-value measures, because the result that there is not added-value is far more conclusive than evidence that there is (i.e., how large the effect size should be for favoring the multiple solution?). Also, more complex procedures for point and interval estimation of all the indices should be considered. In the particular case of the nonlinear-UVA modeling, there is still ample room for further developments, because both reliability and information (and so relative efficiency as well) are conditional here, and what we have proposed are marginal summaries (which is a needed first step). Finally, the impact that the correlated specificities would finally have on external variable relations is an issue that must necessarily be studied in the future (e.g., Reise & Haviland, in press). In particular, it would be interesting to extend the added-value proposal for assessing potential predictive gains. On the positive side, we believe that most of the evidence that is needed in the future should be obtained from real applications, and, in this respect, it is worth noting that the indices and procedures proposed here are intuitive and easy to interpret, and they can be computed using the software that we make freely available for interested researchers and applied professionals.

Declaration of Conflicting Interests

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

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is part of the project I+D+I under grant PID2020-112894GB-I00, funded by MCIN/AEI/10.13039/501100011033, and by the Catalan Ministry of Universities, Research and the Information Society under grant 2021 SGR 00036. The funding source was not involved in any step of the research process, neither in the writing and publication process.

Ethical Approval and Informed Consent

This study was approved by the Research and Innovation Ethics Committee (CEIPSA) of Universitat Rovira i Virgili (CEIPSA-2021-PR-0002). Participants provided written informed consent.

ORCID iDs

Pere J. Ferrando https://orcid.org/0000-0002-3133-5466
Fabia Morales-Vives https://orcid.org/0000-0002-2095-0244

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement*, *20*(4), 311–329. https://doi.org/10.1177/014662169602000402

Bandalos, D. L. (2021). Item meaning and order as causes of correlated residuals in confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 903–913. https://doi.org/10.1080/10705511.2021.1916395

Beauducel, A., & Leue, A. (2013). Unit-weighted scales imply models that should be tested. *Practical Assessment, Research & Evaluation*, *18*(1), 1–7. https://doi.org/10.7275/y3cg-xv71

Bollen, K. A., & Stine, R. A. (1992). Bootstrapping goodness-of-fit measures in structural equation models. *Sociological Methods & Research*, *21*(2), 205–229. https://doi.org/10.1177/0049124192021002004

Calderon, C., Navarro-Gonzalez, D., Lorenzo-Seva, U., & Ferrando, P. J. (2019). Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for assessing the dimensionality of tests and items. *Psicothema*, *31*(4), 450–457. https://doi.org/10.7334/psicothema2019.153

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2024). *shiny: Web application framework for R* (R Package Version 1.9.1.9000). https://github.com/rstudio/shiny

Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. https://doi.org/10.1111/1467-8721.ep10768783

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.

Cronbach, L. J., & Gleser, G. C. (1964). The signal/noise ratio in the comparison of reliability coefficients. *Educational and Psychological Measurement*, *24*(3), 467–480. https://doi.org/10.1177/001316446402400303

DeMars, C. E. (2020). Comparing causes of dependency: Shared latent trait or dependence on observed response. *Journal of Applied Measurement*, *21*(4), 400–419. https://europepmc.org/article/med/33989197

Drasgow, F., & Miller, H. E. (1982). Psychometric and substantive issues in scale construction and validation. *Journal of Applied Psychology*, *67*(3), 268–279. https://doi.org/10.1037/0021-9010.67.3.268

Dueñas, J. M., Morales-Vives, F., Camarero-Figuerola, M., & Tierno-García, J. M. (2022). Spanish adaptation of the Family Involvement Questionnaire-High School: Version for parents. *Psicología Educativa*, *28*(1), 31–38. https://doi.org/10.5093/psed2020a21

Edwards, M. C., Houts, C. R., & Cai, L. (2018). A diagnostic procedure to detect departures from local independence in item response theory models. *Psychological Methods*, *23*(1), 138–149. https://doi.org/10.1037/met0000121

Ferrando, P. J., & Lorenzo-Seva, U. (2019). On the added value of multiple factor score estimates in essentially unidimensional models. *Educational and Psychological Measurement*, *79*(2), 249–271. https://doi.org/10.1177/0013164418773851

Ferrando, P. J., & Morales-Vives, F. (2023). Is it quality, is it redundancy, or is model inadequacy? Some strategies for judging the appropriateness of high-discrimination items. *Anales de Psicología*, *39*(3), 517–527. https://doi.org/10.6018/analesps.535781

Ferrando, P. J., Navarro-González, D., & Lorenzo-Seva, U. (2024). A relative normed effect-size difference index for determining the number of common factors in exploratory solutions. *Educational and Psychological Measurement*, *84*(4), 736–752.

Ferrando, P. J., Navarro-Gonzalez, D., & Morales-Vives, F. (2025). Linear and nonlinear indices of score accuracy and item effectiveness for measures that contain locally

dependent items. *Educational and Psychological Measurement*, *85*(1), 60–81. https://doi.org/10.1177/00131644241257602

Furnham, A. (1990). The development of single trait personality theories. *Personality and Individual Differences*, *11*(9), 923–929. https://doi.org/10.1016/0191-8869(90)90273-T

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika*, *75*, 209–227. https://doi.org/10.1007/s11336-010-9158-4

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Lawrence Erlbaum Associates.

Lucke, J. F. (2005). ''Rassling the hog'': The influence of correlated item error on internal consistency, classical reliability, and congeneric reliability. *Applied Psychological Measurement*, *29*(2), 106–125. https://doi.org/10.1177/0146621604272739

Mansolf, M., & Reise, S. P. (2018). Case diagnostics for factor analysis of ordered categorical data with applications to person-fit measurement. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*(1), 86–100. https://doi.org/10.1080/10705511.2017.1367926

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology*, *34*(1), 100–117. https://doi.org/10.1111/j.2044-8317.1981.tb00621.x

McDonald, R. P. (1985). *Factor analysis and related methods*. Psychology Press. https://doi.org/10.4324/9781315802510

McDonald, R. P. (1999). *Test theory: A unified approach*. Routledge.

McDonald, R. P. (2000). A basis for multidimensional item response theory. *Applied Psychological Measurement*, *24*(2), 99–114. https://doi.org/10.1177/01466210022031552

Mulaik, S. A. (2010). *Foundations of factor analysis* (2nd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b15851

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, *49*(1), 115–132. https://doi.org/10.1007/BF02294210

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. Bollen & S. J. Long (Eds.), *Testing structural equation models* (pp. 205–243). Sage.

Muthén, L. K., & Muthén, B. (2017). *Mplus: Statistical analysis with latent variables: User's guide* (Version 8).

Nicewander, W. A. (1993). Some relationships between the information function of IRT and the signal/noise ratio and reliability coefficient of classical test theory. *Psychometrika*, *58*, 139–141. https://doi.org/10.1007/BF02294477

Raykov, T. (2001). Estimation of congeneric scale reliability using covariance structure analysis with nonlinear constraints. *British Journal of Mathematical and Statistical Psychology*, *54*(2), 315–323. https://doi.org/10.1348/000711001159582

Raykov, T., & Calvocoressi, L. (2021). Model selection and average proportion explained variance in exploratory factor analysis. *Educational and Psychological Measurement*, *81*(6), 1203–1220. https://doi.org/10.1177/0013164420963162

Raykov, T., Gabler, S., & Dimitrov, D. M. (2015). Maximal reliability and composite reliability: A latent variable modeling approach to their difference evaluation. *Structural Equation Modeling*, *23*(3), 384–391. https://doi.org/10.1080/10705511.2014.966369

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Raykov, T., & Marcoulides, G. A. (2015). On examining the underlying normal variable assumption in latent variable models with categorical indicators. *Structural Equation Modeling: A Multidisciplinary Journal*, *22*(4), 581–587.

Reise, S. P., & Haviland, M. G. (in press). Understanding alpha and beta and sources of common variance: Theoretical underpinnings and a practical example. *Journal of Personality Assessment*. https://doi.org/10.1080/00223891.2024.2420175

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment*, *92*(6), 544–559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27–48. https://doi.org/10.1146/annurev.clinpsy.032408.153553

Revelle, W. (1979). Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research*, *14*(1), 57–74. https://doi.org/10.1207/s15327906mbr1401_4

Revelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*, 145–154. https://doi.org/10.1007/s11336-008-9102-z

Rindskopf, D., & Rose, T. (1988). Some theory and applications of confirmatory second-order factor analysis. *Multivariate Behavioral Research*, *23*(1), 51–67. https://doi.org/10.1207/s15327906mbr2301_3

Schweizer, K., Gold, A., Krampen, D., & Troche, S. (2024). Conceptualizing correlated residuals as item-level method effects in confirmatory factor analysis. *Educational and Psychological Measurement*, *84*(5), 869–886.

Sijtsma, K., Ellis, J. L., & Borsboom, D. (2024). Recognize the value of the sum score, psychometrics' greatest accomplishment. *Psychometrika*, *89*(1), 84–117. https://doi.org/10.1007/s11336-024-09964-7

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin*, *83*(2), 213. https://doi.org/10.1037/0033-2909.83.2.213

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22–29. https://doi.org/10.1111/j.1745-3992.1996.tb00803.x

Widaman, K. F., & Revelle, W. (2023). Thinking thrice about sum scores, and then some more about measurement and analysis. *Behavior Research Methods*, *55*(2), 788–806. https://doi.org/10.3758/s13428-022-01849-w

Zinbarg, R. E., Revelle, W., & Yovel, I. (2007). Estimating ω h for structures containing two group factors: Perils and prospects. *Applied Psychological Measurement*, *31*(2), 135–157. https://doi.org/10.1177/0146621606291558