Psicothema

Psicothema 2019, Vol. 31, No. 4, 450-457
doi: 10.7334/psicothema2019.153

ISSN 0214 - 9915 CODEN PSOTEG

www.psicothema.com

Multidimensional or essentially unidimensional? A multi-faceted factor-analytic approach for assessing the dimensionality of tests and items

Caterina Calderón Garrido[1], David Navarro González[2], Urbano Lorenzo Seva[2], and Pere J. Ferrando Piera[2]
[1] Universidad de Barcelona and [2] Universidad Rovira i Virgili

Abstract

Background: Factor-analysis based dimensional assessment of psychometric measures is a key step in the development of tests. However, current practices for deciding between a multiple-correlated or essentially unidimensional solution are clearly improvable. **Method:** A series of recent studies are reviewed, and an approach is proposed that combines multiple sources of information, which is expected to be used to make an informed judgement about the most appropriate dimensionality for the measure being studied. It uses both internal and external sources of information, and focuses on the properties of the scores derived from each of the solutions compared. **Results**: The proposal is applied to a re-analysis of a measure of symptoms of psychological distress. The results show that a clear and informed judgement about the most appropriate dimensionality of the measure in the target population can be obtained. **Discussion:** The proposal is useful and can be put into practice by using user-friendly, non-commercial software. We hope that this availability will result in good practice in the future.

Keywords: Exploratory factor analysis, essential unidimensionality factor score estimates, marginal reliability, external validity.

Resumen

¿Multidimensional o esencialmente unidimensional? Una propuesta multifacética basada en el análisis factorial para evaluar la dimensionalidad de ítems y tests. **Antecedentes:** la evaluación de la dimensionalidad de las medidas psicométricas mediante análisis factorial es un aspecto básico en la construcción y desarrollo de tests. Sin embargo, las prácticas utilizadas habitualmente para decidir entre soluciones múltiples o esencialmente unidimensionales son bastante mejorables. **Método**: se revisan una serie de trabajos recientes y se propone una aproximación basada en múltiples fuentes de información, que permite tomar decisiones informadas acerca de la dimensionalidad más apropiada para la medida que se evalúa. La propuesta utiliza tanto fuentes internas como externas y se basa sobre todo en las propiedades de las puntuaciones que se derivan de cada una de las soluciones a comparar. **Resultados:** la propuesta se aplica como ejemplo ilustrativo en un re-análisis de una medida de síntomas referidos al malestar psicológico. Los resultados muestran que es posible tomar una decisión clara acerca de la dimensionalidad más apropiada de la medida en la población de referencia. **Conclusión:** la propuesta se considera útil y además puede llevarse a cabo mediante el uso de programas no comerciales. Se espera que esta disponibilidad pueda llevar al uso de mejores prácticas en el futuro.

Palabras clave: análisis factorial exploratorio, unidimensionalidad esencial, puntuaciones factoriales estimadas, fiabilidad marginal, validez externa.

Determining the number of dimensions that underlie item scores is possibly the most basic issue in the assessment of the psychometric properties of a psychometric instrument, and factor analysis (FA) is by far the most widely used tool for addressing this issue (Izquierdo, Olea, & Abad, 2015, Mershon & Gorsuch, 1988, Muñiz & Fonseca-Pedrero, 2019). Furthermore, the criteria for judging dimensionality has evolved, as expected, almost in parallel to how FA itself has evolved. Thus, since the 1970s, judgements have increasingly been based on the values of the goodness-of-fit indices derived from fitting a model with the specified number of factors to the data. In recent years, however, the tide seems to be turning (Ferrando & Lorenzo-Seva, 2017).

Most psychometric measures, especially in the personality domain, were initially designed to measure a single construct (Furnham, 1990). However, virtually all the item scores derived from these measures fail to meet the strict goodness-of-fit criteria of unidimensionality required by the single-factor FA model. When this occurs, the predictable next move is to fit multiple correlated FA solutions to the data and propose the resulting solutions (which are better fitting) as the most appropriate structures for the measures under scrutiny (Ferrando & Lorenzo-Seva, 2018a, 2018b, Furnham, 1990, Reise, Bonifay, & Haviland, 2013, Reise, Cook, & Moore, 2015). However, most instruments designed to be unidimensional do, in fact, yield data that is compatible with a solution in which there is a strong, dominant factor running through all the test items (Floyd & Widaman, 1995, Reise, Bonifay, & Haviland, 2013, Reise, Cook, & Moore, 2015).

Received: May 28, 2019 • Accepted: August 1, 2019
Corresponding author: Pere J. Ferrando Piera
Universidad Rovira i Virgili
43007 Tarragona (Spain)
e-mail: perejoan.ferrando@urv.cat

As the idea that dimensionality cannot be appraised on the sole basis of goodness of model-data fit has been gaining momentum, psychometricians have started to propose alternative or complementary approaches (Ferrando & Lorenzo-Seva, 2018a, 2018b, Raykov & Marcoulides, 2018, Rodriguez, Reise, & Haviland, 2016a, 2016b). The basis of these proposals can be summarized in two points. First, the assessment must be multi-faceted and many aspects of appropriateness must be considered. Second, the point is not whether a set of scores is strictly unidimensional, but rather if it is unidimensional enough so that (a) the item parameter estimates are unbiased, (b) reliable and valid inferences can be made from scores based on the single-factor solution, and (c) no essential information is lost when this solution is adopted (Reise, Bonifay, & Haviland, 2013).

Point (b) above leads to a third change of focus, which we firmly believe in. The ultimate purpose of most measuring instruments is individual measurement in some form (e.g. Cliff, 1977). So, when dimensionality is being judged, the emphasis should not be (or not totally be) on the goodness of fit and factor structure of the solution, but rather on the properties of the score estimates derived from this solution (Ferrando & Lorenzo-Seva, 2018b, Ferrando & Navarro-González, 2018).

The consequences of adopting a wrong decision can now be considered from the views above. If item scores are essentially unidimensional but are treated as multidimensional, the main potential consequences are: lack of clarity in the interpretation and unnecessary theoretical complexities; weak, nonreplicable factors of little substantive interest; and (as a consequence) weakened factor score estimates that do not allow accurate individual measurements to be made. On the other hand, treating clearly multidimensional scores as unidimensional is expected to lead to biased item parameter estimates, loss of information, and factor score estimates that cannot be univocally interpreted because they reflect the impact of multiple sources of variance (see Ferrando & Navarro-González, 2018, Reise, Bonifay, & Haviland, 2013, and Reise, Cook, & Moore, 2015).

The present article integrates and summarizes several procedures, designed to assess dimensionality from an FA perspective, that have been proposed in a series of papers over the last three years. The resulting approach is comprehensive and multifaceted and uses internal information from the item scores and external information via relations to related variables. As discussed above, it is also heavily oriented towards the properties of the factor score estimates derived from a given solution. No claim is made for new methodological developments. So, the aim of this article is mainly illustrative and of substantive interest.

The proposed approach

Consider two competing FA solutions to be compared: a multiple correlated solution, and a unidimensional solution. The latter can be obtained either by directly fitting the unidimensional FA model to the item scores, or by fitting a second-order solution in which the multiple factors are assumed to measure a common higher-order factor. In both cases we shall denote the resulting factor as the general factor. Furthermore, by using second-order terminology, we shall denote the multiple factors of the first solution as primary factors.

The FA solutions to be compared can be based either on the linear FA model or on the IRT-related non-linear model (e.g. Ferrando & Lorenzo-Seva, 2013, Suàrez-Alvarez et al. 2018, Villegas et al. 2018). In principle, the procedures could also be applied to both unrestricted and restricted solutions. However, given the basic nature of the assessment, the proposal makes much more sense in the context of unrestricted or exploratory solutions. With regards to estimation procedures, the first-stage item calibration can be based on any of the existing criteria for fitting the FA model. As for the second stage (scoring), the procedures available to date have been developed for two types of factor score estimates: maximum likelihood (ML; Bartlett scores in the linear case), and expected-a-posteriori (EAP; Regression scores in the linear case).

For didactic and illustrative purposes, the approach is described by using a sequence of three stages. Applications do not need to adhere to this sequential approach, but it describes the most natural order in which the results of the assessment can be presented.

First stage: Basic internal assessment

The first step in this stage is conventional goodness-of-fit (GOF) assessment. This is the usual approach for judging dimensionality, so we shall not discuss it in any further detail (see e.g. Suàrez-Alvarez et al. 2018 and Villegas et al. 2018)

The second step is to assess the degree of dominance of the general factor or closeness to unidimensionality. A simple and informative index for doing so is the explained common variance (ECV) index (e.g. ten Berge & Kiers, 1991), which essentially measures the proportion of common variance of the item scores that can be accounted for by the first canonical factor (i.e. the factor that explains most common variance). We shall stress here that this index differs from the usual criterion of the amount of **total** variance accounted for by the first factor (see Ferrando & Lorenzo-Seva, 2017). As for interpretation, the more common variance the first canonical factor explains: (a) the closer to unidimensionality the data is, (b) the less likely it is that the factor structure based on a single general factor will be biased by multidimensionality effects, and (c) the more univocally the factor score estimates derived from the unidimensional solution can be interpreted. As for reference values, it has been proposed that ECV values should be in the range .70 to .85 if it is to be concluded that a solution can be treated as essentially unidimensional (Rodriguez et al., 2016a, 2016b).

The final step at this stage is about the quality and properties of the factor score estimates derived from the factor solutions. Regardless of which of the competing solutions is judged as the most appropriate, if the derived score estimates do not attain a minimal degree of quality, then the solution cannot be considered of interest from a measurement perspective.

The starting point for assessing the properties of the score estimates is that consistency of person ordering is the primary goal of individual assessment (Cliff, 1977). So, the quality of the factor score estimates should be judged by the extent to which they can consistently order the respondents along the factor continuum. Now, a basic index for addressing this property is the (marginal) reliability of the factor score estimates, denoted here as $\rho^2_{(\hat{\theta}\hat{\theta})}$. By using one of the standard definitions of reliability (Lord & Novick, 1969) $\rho^2_{(\hat{\theta}\hat{\theta})}$ can be interpreted as the squared correlation between the factor score estimates and the levels on the factors they estimate. So, when $\rho^2_{(\hat{\theta}\hat{\theta})}$ is high, the factor score estimates are: (a) accurate with low measurement error; (b) good proxies for representing the true factor scores; and (c) effective for differentiating between respondents with different trait levels (Ferrando & Navarro-González, 2018, Ferrando, Navarro-González, & Lorenzo-Seva, in press). As for

reference values, .80 is a reasonable minimal requirement if the score estimates are to be used for individual assessment (Ferrando & Lorenzo-Seva 2018b, Rodriguez et al., 2016a).

The second index proposed at the score level is the "expected percentage of true differences" (EPTD; Ferrando, Navarro-González, & Lorenzo-Seva, in press). EPTD is the population-expected percentage of differences between the factor score estimates that are in the same direction as the corresponding true differences. So, EPTD assesses one aspect of the "consistency of person ordering" property which is different from (although related to) the one assessed by marginal reliability: that's to say, not the size of the differences that can be detected (reliability), but the proportion of differences (of any size) expected to reflect true differences in the same direction. As for reference values, 50% would indicate that any differences among factor score estimates are random and that this value cannot be used to differentiate or order individuals. Ferrando, Navarro-González, & Lorenzo-Seva (in press) suggested a minimum EPTD cut-off value of 90% if factor score estimates are to be used for individual assessment.

Second Stage: Added-Value assessment

The added-value principle was initially proposed by Haberman (2008) in the context of sub-scale scores. Adapted to the present scenario, the idea is to assess whether the factor score estimates from a primary factor are more accurate predictors of the corresponding primary true scores than the score estimates from the general factor are (Ferrando & Lorenzo-Seva, 2018b). Although the opposite result seems counterintuitive, it is possible, and is expected to occur when: (a) the primary factors are highly correlated with one another, and (b) the general factor score estimates are far more reliable than the primary factor score estimates.

The implications of the idea above for the present purposes are clear. When the primary estimates can more accurately predict their corresponding factors than the general estimates can, they are considered to have "added value", and this result would indicate that the multidimensional solution is the most appropriate. If, on the contrary, the primary factors can be better predicted from the general estimates, then, the multidimensional model is not expected to provide useful information beyond that provided by the unidimensional model, and the choice of the most complex model has little justification.

Operatively, the index for judging whether the primary estimates have added value or not is the "proportional reduction of mean squared error" (PRMSE, see Ferrando & Lorenzo-Seva, 2018b). The larger the reduction, the more accurate the prediction is. So, for each primary factor, the added-value assessment simply compares the PRMSE obtained from the corresponding estimates to the PRMSE obtained from the general estimates, and considers that there is added value if $PRMSE_k > PRMSE_g$. A more refined strategy which is available at present (see below) is to obtain a confidence interval for $PRMSE_k$ and assess whether the lower limit of this interval is still above $PRMSE_g$.

Third Stage: external validity assessment

In the FA literature several authors (e.g. Floyd & Widaman, 1995, Mershon & Gorsuch, 1988) have stated that internal evidence should not be the main criterion for judging the appropriateness of a solution. Rather, the ultimate criterion should be based on 'outside' validity evidence: how the factors in the solution relate to relevant external variables or criteria. Ferrando & Lorenzo-Seva (2019) proposed an integrated approach which is based on this idea. For simplicity, we shall denote this procedure as UNIVAL, which is the name of the package that implements it.

The null model on which UNIVAL is based is that there is a general common factor running throughout the set of items and that all the relations between the external variables and the primary factors in the multiple solution are mediated by this general factor. This is a second-order schema in which the primary factors do not provide further validity information beyond that which can be obtained from the general factor. If this model holds, then, the most parsimonious unidimensional solution has to be considered as the most appropriate in validity terms.

The validity relations between the factor score estimates and the external variables are next assessed on the basis of the null expectations above. Two facets of validity – differential and incremental – are considered. In differential validity terms, what is expected from the null model is that the primary score estimates are related to the external variable in the same way as they relate to the general factor. As for incremental validity, the expectation is that the prediction of the external variable which is made from the general factor score estimates cannot be improved upon by using the primary factor score estimates in a multiple regression schema. When the null model does not hold at all, however, then both information and prediction accuracy is lost if the unidimensional model is chosen in place of the multiple model. And, if the loss is substantial, the multiple model will be the best choice in validity terms.

Operatively, UNIVAL corrects the primary score estimates for measurement error. With regards to differential validity, the null model implies that the disattenuated correlations between the primary score estimates and the external variables are proportional to the corresponding correlations between the primary factors and the general factor. Ferrando & Lorenzo-Seva (2019) proposed scaling the disattenuated correlations by using this proportionality result, so the scaled coefficients are expected to have the same value when the null model holds. A simple approach for assessing if this is so is to obtain confidence intervals for the scaled coefficients and check whether they overlap or not.

As for incremental validity, Ferrando & Lorenzo-Seva (2019) derived corrected single (denoted by $\rho_{c\hat{\theta}_g}$) and multiple (denoted by R_c) correlation coefficients expected to have the same value under the null model. On the other hand, under the alternative model, R_c should always be larger. In line with the incremental validity concept (Raykov & Marcoulides, 2018), the test here is to compute the ($R_c - \rho_{c\hat{\theta}_g}$) difference, set a confidence interval for it, and check whether the zero value lies within the interval (i.e. no incremental validity) or not (incremental validity).

Implementation

The authors' experience suggests that proposals such as the present one are only used in applications if they are implemented in available software. In this respect, the two first stages have been implemented in version 10.9 of the program FACTOR (Ferrando & Lorenzo-Seva, 2017). In order to compute them with FACTOR, researchers have to specify the number of group factors. In addition, in the "Other specifications of factor model" menu they have to check the options: (a) "Closeness to unidimensional assessment"; (b) "Added value of multiple score estimates: first

(or second) order factor model"; and (c) "Assess quality of factor scores". An output containing the info relative to these indices is presented in Figure 1.

The third stage can be implemented in R software using the UNIVAL package available through CRAN (https://CRAN.R-project.org/package=unival). The function works for both ML and EAP scores and also for both linear and non-linear FA models.

The main function of the package is also named **unival**, and, using the minimal input arguments, is executed by using the following R command:

```
unival(y, FP, fg, PHI, FA_model = 'Linear', type)
```

where the arguments are the following:

y: Related external variable.

FP: Primary factor score estimates.

fg: General or second-order factor score estimates.

PHI: Inter-Factor correlation matrix.

FA_model: Which FA-model was used for calibration and scoring. Available options

are: "Linear" (by default) or "Graded".

type: Which type of factor score estimates were used in FP and fg. The two available options are: "ML" or "EAP" scores. If not specified, ML will be assumed.

```
CLOSENESS TO UNIDIMENSIONALITY ASSESSMENT
Ferrando & Lorenzo-Seva (2018)

[...]

OVERALL ASSESSMENT

UniCo  =   0.928
ECV    =   0.850
MIREAL =   0.258

------------------------------------------------------------------------------

[...]

------------------------------------------------------------------------------

QUALITY AND EFFECTIVENESS OF FACTOR SCORE ESTIMATES

Ferrando & Lorenzo-Seva (2018)

                                              F  1      F  2      F  3

Factor Determinacy Index (FDI)                0.976     0.959     0.922
ORION marginal reliability                    0.952     0.919     0.850
Sensitivity ratio (SR)                        4.456     3.366     2.376
Expected percentage of true differences (EPTD)  95.7%     92.3%     90.4%

The sensitivity ratio (SR) can be interpreted as the number of different factor levels than can be differentiated
on the basis of the factor score estimates. The expected percentage of true differences (EPTD) is the estimated
percentage of differences between the observed factor score estimates that are in the same direction as the
corresponding true differences.

If factor scores are to be used for individual assessment, FDI values above .90, marginal reliabilities above .80,
SR above 2, and EPTDs above 90% are recommended.


------------------------------------------------------------------------------

ADDED VALUE OF MULTIPLE FACTOR SCORE ESTIMATES: PROPORTIONAL REDUCTION IN MEAN SQUARED ERROR OF PREDICTION (PRMSE)
Ferrando & Lorenzo-Seva (2019)

Factor model assessed: a first order general factor versus a multiple factor model (empirical based MSE)

Factor     From Primary Factors        From the General Factor   From both Primary Factors and
                                                                 the General Factor
F   1      0.952                       0.699                     0.975
F   2      0.919                       0.785                     0.958
F   3      0.850                       0.446                     0.919
```

Figure 1. Output of the selected specifications of the model in FACTOR

The outcome includes the differential and incremental validity coefficients along with their corresponding bootstrap confidence intervals. The results can be printed on console and/or saved as new variables.

The package contains a considerable amount of documentation with more information about the input and output arguments, the usability of the procedure, and an example of implementation. The documentation is also accessible through CRAN (https://cran.r-project.org/web/packages/unival/unival.pdf).

Method

Participants

In recent years, NEOcoping, a national, multicenter, cross-sectional, prospective study by the Continuous Care Group of the Spanish Society of Medical Oncology (SEOM) has been interested in assessing the prevalence of psychological distress in cancer patients. NEOcoping researchers considered that the BSI-18 was potentially useful for assessing this issue. However, if BSI-18 is to be used, the most appropriate dimensionality of the scores in the target population should first be determined. The present example is a re-analysis of some of the data presented in Calderon et al., (2019). The sample used for assessing dimensionality consisted of 877 patients with histologically confirmed, non-advanced cancer (see Calderon et al., (2019), for more details).

Instruments

The *Brief Symptom Inventory-18* (BSI-18, Derogatis, 2001), is an extremely popular instrument for assessing psychological distress. It is made up of 18 items which ask the respondents how they have felt the last 7 days. Each item is rated on a 5-point Likert scale from 0 (not at all) to 4 (extremely), and the items are organized into three subscales of 6 items each: Depression (DEP), Anxiety (ANX) and Somatizations (SOM). Raw scores can be obtained either at the subscale level or at the total level, which is known as the Global Severity Index (GSI).

As expected, there are many FAs on the most appropriate structure for the BSI-18, and most of them try to decide between a tri-dimensional solution, with three strongly correlated factors, and a unidimensional solution (see Calderon et al., 2019, for a review). Also as expected, the results are far from conclusive.

The Multidimensional Scale of Perceived Social Support (MSPSS; Zimet, Powell, Farley, Werkman, & Berkoff, 1990) was also administered to the sample above. The MSPSS is a 12-item self-report that measures the perceptions and adequacy of social support from three sources: family, friends, and significant others. It also allows a total score to be obtained that reflects the overall degree of social support that the patients receive. Negative relations are expected between psychological distress, as measured with the BSI-18, and social support, as measured by the MSPSS (Calderon et al,. 2019). So, Family, Friends, Significant-Others, and total MSPSS scores were used as external variables in the third-stage assessment.

Procedure

Participants were asked to answer to the BSI-18 and the MSPSS according to the conditions described in Calderon et al. (2019).

Data analysis

Analysis were conducted by using FACTOR 10.9 (Ferrando & Lorenzo-Seva, 2017) and the UNIVAL package (Ferrando, Lorenzo-Seva, & Navarro-González, in press), available through R.

Results

First-Stage Results

The descriptive statistics showed that item scores tended to be asymmetrical (positively skewed), some of them strongly so. Taking this into account and the fact that the test is not too long and the sample is reasonably large, the best model to fit the data is nonlinear FA. In this model, the item scores are treated as ordered-categorical variables and the FA is fitted to the inter-item polychoric correlation matrix (see Ferrando & Lorenzo-Seva, 2013). The chosen fitting function was robust unweighted least squares, with mean-and-variance corrected fit statistics (Ferrando & Lorenzo-Seva, 2017). As preliminary analyses, sampling adequacy was assessed by the KMO and Bartlett's sphericity test, and was considered to be very good.

Models with between 1 and 3 factors were fitted to the data. In accordance with previous results, the present ones suggested that the decision was between the unidimensional and the three-factor. GOF results for these two solutions as well the general factor dominance results are in the upper panel of table 1.

In summary, the fit of the unidimensional model is only marginally acceptable in pure GOF terms, whereas the fit of the model in 3 factors is excellent by all the standards. However, the ECV value suggests that there is a strong dominant factor running through all the 18 items, and the PA-based procedure (Timmerman & Lorenzo-Seva, 2011) suggests that the unidimensional solution is the most replicable.

To obtain further information at the structural level, first the solution in three factors was rotated to achieve maximum factor simplicity by using the Promin criterion (Lorenzo-Seva, 1999). The rotated pattern closely approached simple structure (Bentler's simplicity index was .98) and allocated all the items in the

Table 1
Stage-1 results: Basic internal assessment

Panel (a)

Model	RMSEA	CFI	ECV	PA
1 factor	0.073	0.981	0.850	1 factor (62.37%)
3 factors	0.032	0.997	–	–

Note: RMSEA=Root Mean Square Error of Approximation; CFI=Comparative Fit Index; ECV= Explained Common Variance; PA: Parallel Analysis

Panel (b)

	Marginal reliability	EPDT
General factor	0.955	95.8%
DEP factor	0.919	92.3%
ANX factor	0.952	95.7%
SOM factor	0.850	90.4%

expected 'a priori' structure. Second, a second-order solution with a single general factor was obtained based on the primary inter-factor correlation matrix. The loadings of the items on the second-order factor were quite similar to those obtained by directly fitting the unidimensional model to the data (the coefficient of factor congruence was .88; see Lorenzo-Seva & ten Berge, 2006). Furthermore, the general-factor structure had positive-manifold, with all the loadings above .30. These results provide additional support to the hypothesis that there is a general, dominant factor running through all 18 items (see Ferrando & Lorenzo-Seva, 2019, Mershon & Gorsuch, 1988). The different solutions discussed so far are available from the authors.

Next, the results of the item calibration were taken as fixed and known, and used to obtain factor score estimates, which were EAP scores. Results about the properties of these scores are in the lower panel of table 1. They are quite clear: The marginal reliabilities of the primary factor score estimates are rather high, which means that they (a) are good proxies for the factors they represent, and (b) allow respondents to be accurately measured in clinical settings. The EPTD values are all above the cut-off value proposed here, which implies that individuals can be consistently ordered and differentiated on the basis of the score estimates. Finally, note that both the marginal reliability and the EPTD of the general factor score estimates are higher than those of any of the three primary factors.

Second-Stage Results

The last result above makes the Added-value assessment quite relevant here. Even when the primary score estimates are quite accurate, the general estimates are even more so. So, it cannot be discarded that better predictions of the primary factors could be obtained on the basis of the general estimates. The results in table 2, however, are quite clear. The three primary factors show added value because, for all of them, the true factor scores are better predicted from the corresponding estimates than from the score estimates in the single general factor. Note, in particular, that the lower limits of the confidence intervals for the primary PRMSEs are above the general estimated PRMSEs. So, accuracy will be lost if the general score estimates are used to predict the 'true' levels of the individuals in the primary factors.

Third-Stage Results

As expected, the relations between all the BSI-18 factor score estimates and the external MSPSS measures of external support were negative. Panel (a) in table 3 shows the corresponding product-moment correlations.

Table 2
Stage-2 results: Added-value assessment

	PRMSE From the primary score estimates	**PRMSE From the general score estimates**
DEP factor	0.919 (0.805;0.973)	0.785
ANX factor	0.952 (0.950;0.965)	0.699
SOM factor	0.850 (0.825;0.861)	0.446

Table 3
Stage-3 results

Factor scores	**γ**	**Family**	**Friends**	**Significant others**	**MSPSS total**
(a) Correlations between BSI-18 factor score estimates and MSPSS scores					
DEP	.980	-.227**	-.167**	-.220**	-.258**
ANX	.846	-.142**	-.091**	-.117**	-.146**
SOM	.696	-.120**	-.123**	-.141**	-.165**
General	-	-.178**	-.136**	-.172**	-.205**

*p<0.01; **p<0.001; MPSS, multidimensional scale of perceived social support, γ, second-order loadings

(b) External-validity assessment

	Differential validity estimates	**Incremental validity estimates**
Family	0.0590 (-0.0039 ; 0.1499)	0.0865 (0.0546 ; 0.1715)
Friends	0.0131 (-0.1162 ; 0.1359)	0.0785 (0.0290 ; 0.1635)
Sig. Others	0.0172 (-0.0536 ; 0.1140)	0.1066 (.0712 ; 0.1973)
Total	0.0213 (-0.0512 ; 0.1164)	0.1130 (0.0736 ; 0.2109)

The pattern of correlations in table 3 is quite consistent: the DEP factor score estimates are the ones that are most strongly related to all of the external variables, and the general factor estimates come a second. Also as expected, the strongest relations are found for the total MSPSS scores.

The first column in panel (a) shows the second-order loadings of the primary factors on the general factor. Note that the DEP factor, which is the most related to the external variables is also the most related to the second-order general factor.

Panel (b) in table 3 shows the UNIVAL differential and incremental results. For clarity, all the correlations were reversed to have positive values. Furthermore, for simplicity, the table's first differential column only shows the difference between the most extreme scaled coefficient and the median of all of them. The results are clear here: the zero value falls within the confidence interval in all cases. So, the primary factors seem to relate to all the external variables in essentially the same way as they relate to the general factor. Note that this result could be predicted from the discussion above regarding panel (a).

The incremental results in the second column are also clear but they go in the opposite direction: in all cases, the primary score estimates allow for better predictions of the external variables than those can be obtained from the general score estimates. Overall, these results suggest that some unique parts of the primary factors are still related to the external variables in ways other than those explained by the common general factor. However, these unique parts relate to the external variables essentially in the same way as the corresponding primary factors relate to the general factor.

The dimensionality proposal

The results obtained so far allow us to make an informed judgement about the most appropriate dimensionality of the BSI-18 scores in the target population. Overall, the solution of choice would be the tridimensional one. First, it fits very well, and provides a clear and interpretable structure that agrees with the theoretical design from which the scale was developed. Furthermore, the additional

results from our proposal suggest that it provides accurate score estimates that can be used for individual assessment purposes, and that the standing of the respondent on the primary factors is better predicted from the corresponding estimates than from the general score estimates. Finally, the multiple score estimates provide better predictions of relevant variables of interest than those that can be obtained solely from the general score estimates.

The results also suggest that the BSI-18 scores can be treated as essentially unidimensional at the cost of some loss of information and predictive power. The ECVs, in particular, suggest that there is a strong, dominant factor running through all the test items. So, multidimensionality is not expected to substantially bias the loadings of this solution, and the resulting score estimates can be univocally interpreted as levels on a general dimension of psychological distress. Furthermore, the resulting score estimates are accurate enough to be used in individual assessment. The fine-grained assessment that could be obtained from the tridimensional solution would be lost here, but the single-factor score estimates would possibly be appropriate for a quick or general screening, or for a rank-order of patients in terms of their levels of distress. In validity terms, finally, the predictive power of the general score estimates is significantly worse than that based on the primary estimates. However, the effect sizes for the incremental validities are very small (of about 0.013, at most, in Cohen's f^2 terms). So, the loss of predictive power when general score estimates are used is expected to be very meager in practice.

Discussion

Our experience as counselors and reviewers suggests that, in general, practices for assessing the dimensionality of psychometric measures need to be greatly improved. In this respect, the literature is full of endless rounds of factor analyses and re-analyses of popular measures, which never draw clear conclusions. In our opinions, routine practices and sole reliance on goodness-of-fit and structural assessment are largely to blame for this.

In this article we have proposed a comprehensive approach for FA-based dimensional assessment, which, hopefully, will contribute to encouraging good practices in the applied field. Unlike the conventional way of addressing this issue, we believe that the ultimate aim of most psychometric measures is, precisely, measurement. So, our proposal focuses mostly on the properties of the scores derived from the solution rather than on the structure of the solution itself.

In practical terms, the proposal can be put into practice by using non-commercial software: So, internal stages 1 and 2 can be performed using FACTOR, while the external-validity stage can be performed using UNIVAL. However, we do not know whether this will be sufficient to put an end to routine practices and promote more informed dimensionality assessments in the future. Our proposal requires practitioners to adopt a more active and critical attitude, and the full assessment proposed here requires data from relevant external variables to be collected and used. Muñiz and Fonseca-Pedrero (2019) clearly state that collecting this type of data is an essential step in the development of a psychometric test. However, we do not know if this type of information was available in most of the FA-based studies in the literature.

Acknowledgements

This study was supported by a grant from the Spanish Ministry of Economy and Competitivity (PSI2017-82307-P).

References

Calderon, C., Ferrando, P. J., Lorenzo-Seva, U., Hernández, R., Oporto-Alonso, M., Jiménez-Fonseca, P. (2019). *Testing factor structure and measurement invariance of the BSI-18 across gender, age, tumor site and over time in cancer patients*. Manuscript submitted for publication.

Cliff, N. (1977). A theory of consistency of ordering generalizable to tailored testing. *Psychometrika*, *42*, 375-399. https://doi.org/10.1007/BF02293657

Derogatis, L. R. (2001). *BSI 18, Brief Symptom Inventory 18: Administration, scoring and procedures manual*. Minneapolis: NCS Pearson, Inc.

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

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*, 236-240. https://doi.org/10.7334/psicothema2016.304

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. https://doi.org/10.1177/0013164417719308

Ferrando, P.J. & Lorenzo-Seva, U. (2018b). On the Added Value of Multiple Factor Score Estimates in Essentially Unidimensional Models. *Educational and Psychological Measurement*, *79*(2), 249-271. https://doi.org/10.1177/0013164417719308

Ferrando, P.J. & Lorenzo-Seva, U. (2019). An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement, 79*(3), 437-461. https://doi.org/10.1177/0013164418824755

Ferrando, P.J., Lorenzo-Seva, U., & Navarro-González, D. (in press). unival: An FA-based R Package For Assessing Essential Unidimensionality Using External Validity Information. *R Journal*.

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. https://doi.org/10.1016/j.paid.2017.11.014

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

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment*, *7*, 286-299. https://doi.org/10.1037/1040-3590.7.3.286

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

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204-229. https://doi.org/10.3102/1076998607302636

Izquierdo, I., Olea, J., & Abad, F. J. (2014). Exploratory factor analysis in validation studies: Uses and recommendations. *Psicothema, 26*, 395-400. https://doi.org/10.7334/psicothema2013.349

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

Lorenzo-Seva, U. (1999). Promin: A method for oblique factor rotation. *Multivariate Behavioral Research*, *34*, 347-356. https://doi.org/10.1207/S15327906MBR3403_3

Lorenzo-Seva, U., & Ten Berge, J. M. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology*, *2*, 57-64. https://doi.org/10.1027/1614-2241.2.2.57

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. https://doi.org/10.1037/0022-3514.55.4.675

Muñiz, J., & Fonseca-Pedrero, E. (2019). Diez pasos para la construcción de un test [Ten steps for test development]. *Psicothema, 31*(1), 7-16. https://doi.org/10.7334/psicothema2018.291

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. https://doi.org/10.1177/0013164416678650

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. https://doi.org/10.1080/00223891.2012.725437

Reise, S. P., Cook, K. F., & Moore, T. M. (2015). Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. *Handbook of item response theory modeling. New York: Routledge*, 13-40.

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016a). Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, *21*, 137. https://doi.org/10.1037/met0000045

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016b). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of personality assessment*, *98*, 223-237. https://doi.org/10.1080/00223891.2015.1089249

Suárez-Alvarez, J., Pedrosa, I., Lozano Fernández, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30,* 149-158. https://doi.org/10.7334/psicothema2018.33

Ten Berge, J. M., & Kiers, H. A. (1991). A numerical approach to the approximate and the exact minimum rank of a covariance matrix. *Psychometrika*, *56*, 309-315. https://doi.org/10.1007/BF02294464

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality Assessment of Ordered Polytomous Items with Parallel Analysis. *Psychological Methods, 16*, 209-220. https://doi.org/10.1037/a0023353

Villegas, G., González, N. , Sanchez-García, A. B., Sánchez, M., & Galindo-Villardón, M. P. (2018). Seven methods to determine the dimensionality of tests: Application to the General Self-Efficacy Scale in twenty-six countries. *Psicothema*, *30*, 442-448. https://doi.org/10.7334/psicothema2018.113

Zimet, G. D., Powell, S. S., Farley, G. K., Werkman, S., & Berkoff, K. A. (1990). Psychometric characteristics of the Multidimensional Scale of Perceived Social Support. *Journal of Personality Assessment*, *55*(3-4), 610-617. https://doi.org/10.1080/00223891.1990.9674095