# unival: An FA-based R Package For Assessing Essential Unidimensionality Using External Validity Information

*by Pere J. Ferrando, Urbano Lorenzo-Seva and David Navarro-Gonzalez*

**Abstract**

The **unival** package is designed to help researchers decide between unidimensional and correlated-factors solutions in the factor analysis of psychometric measures. The novelty of the approach is its use of *external* information, in which multiple factor scores and general factor scores are related to relevant external variables or criteria. The **unival** package's implementation comes from a series of procedures put forward by Ferrando and Lorenzo-Seva (2019) and new methodological developments proposed in this article. We assess models fitted using **unival** by means of a simulation study extending the results obtained in the original proposal. Its usefulness is also assessed through a real-world data example. Based on these results, we conclude **unival** is a valuable tool for use in applications in which the dimensionality of an item set is to be assessed.

## Introduction

Assessing the dimensionality of a set of items is one of the central purposes of psychometric factor analysis (FA) applications. At present, both the exploratory (EFA) and the confirmatory (CFA) models can be considered to be fully developed structural equation models (Ferrando and Lorenzo-Seva, 2017). So, in principle, dimensionality can be rigorously assessed by using the wide array of goodness-of-fit procedures available for structural models in general. However, it is becoming increasingly clear that reliance on goodness-of-fit alone is not the way to judge the most appropriate dimensionality for studying a particular set of item scores (Rodriguez et al., 2016a,b).

The problem noted above is particularly noticeable in instruments designed to measure a single trait. In the vast majority of cases, item scores derived from these instruments fail to meet the strict unidimensionality criteria required by Spearman's model. This failure, in turn, led to the proposal of multiple correlated-factor solutions as the most appropriate structure for them (Ferrando and Lorenzo-Seva, 2018, in press; Furnham, 1990; Reise et al., 2013, 2015). However, most instruments designed to be unidimensional do, in fact, yield data compatible with an essentially unidimensional solution (Floyd and Widaman, 1995; Reise et al., 2013, 2015). When this is the case, treating the item scores as multidimensional has several undesirable consequences, mainly, (a) lack of clarity in the interpretation and unnecessary theoretical complexities, and (b) weakened factor score estimates that do not allow accurate individual measurement (Ferrando and Lorenzo-Seva, 2018, in press; Furnham, 1990; Reise et al., 2013, 2015). Indeed, treating clearly multidimensional scores as unidimensional also has such negative consequences as biased item parameter estimates, loss of information, and factor score estimates that cannot be univocally interpreted (see Ferrando and Lorenzo-Seva, 2018; Reise et al., 2013).

In recent years, several indices and criteria have been proposed to assess dimensionality using different perspectives of model appropriateness. These developments, in turn, have been integrated in comprehensive proposals addressing the dimensionality issue from multi-faceted views including, but are not limited to, standard goodness-of-fit results (Ferrando and Lorenzo-Seva, 2018; Raykov and Marcoulides, 2018; Rodriguez et al., 2016a,b). It is worth noting these approaches generally reflect a trend in which the measurement part of the FA model is again relevant (e.g. Curran et al., 2018). Considering the ultimate aim of most psychometric measures is individual measurement, the scoring stage of the FA should be expected to be the most important part of it (Ferrando and Lorenzo-Seva, 2018, in press). Furthermore, if this view is adopted, a basic criterion for deciding if a given FA solution is appropriate is the extent to which the score estimates derived from this solution are strong, reliable, determinate, unbiased, and clearly interpretable (Ferrando and Lorenzo-Seva, 2018; Beauducel et al., 2016; Furnham, 1990; Reise et al., 2013, 2015). Procedures explicitly based on the quality of the score estimates are already available in widely used programs such as FACTOR (Lorenzo-Seva and Ferrando, 2013), and more sophisticated procedures based on Haberman's (2008) added-value principle have been also proposed (Ferrando and Lorenzo-Seva, in press).

A common characteristic of all the proposals discussed so far is their use of *internal* information from the data exclusively: that is to say, the information provided by the item scores of the measure under study. In contrast, the approach implemented here is based on *external* sources of information: that is to say, the information provided by the relations between the factor score estimates derived from a given solution and relevant external variables or criteria. This additional information is a valuable complementary tool that can help reach a decision on whether the instrument under scrutiny is essentially unidimensional or truly multidimensional.

The present article aims to introduce **unival**, a new contributed R package implementing a recently proposed external procedure of the type described above (Ferrando and Lorenzo-Seva, 2019). It also discusses new methodological developments allowing the procedure to be used in a wider range of situations than those considered in the original proposal. The rest of the article is organized as follows. First, we provide a summary the needed theoretical bases, and explain the new methodological contributions. Then, we give details about the package and how to use it. Finally, we assess the functioning of the program and the new developments proposed here with a simulation study and give real-world data examples.

## Theoretical foundations: A review

Consider two alternative FA solutions – unidimensional and multiple-correlated – which are fitted to a set of item scores. Suppose further that both solutions are found to be acceptable by internal criteria, a situation which is quite usual in applications (e.g. Ferrando and Navarro-Gonzalez, 2018). The aim of the proposal summarized here is to assess which of the competing solutions is more appropriate in terms of external validity.

The null hypothesis in the proposal assumes (a) there is a general common factor running through the entire set of items, and (b) all the relations between the multiple factors and the relevant external variables are mediated by the general factor. In this case, the unidimensional solution is the most appropriate in terms of validity. At this point we note the proposal is intended to work on a variable-by-variable basis. So, it will be summarized using a single external variable.

The null hypothesis above can be described by using a second-order FA schema as follows. Assumption (a) above implies the correlated factors in the multiple solution, which we shall denote from now on as primary factors, behave as indicators of a single general factor. Assumption (b) implies the only parts of the primary factor not accounted for by the general factor are unrelated to the external variable.

The implications of the null model in terms of validity relations are considered in two facets: differential and incremental. In differential validity terms, the score estimates derived from the primary factors are expected to be related to the external variable in the same way as they are related to the general factor. As for incremental validity, the implications of the null model are the prediction of the external variable which is made from the single (general) factor score estimates cannot be improved upon by using the primary factor score estimates in a multiple regression schema.

Let $\hat{\theta}_{ik}$ be the factor-score estimate of individual $i$ in the $k$ primary factor, and let $\theta_{ik}$ be the corresponding true factor score. We write

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

where $\varepsilon_{ik}$ denotes the measurement error. The true scores $\theta_k$ are assumed to be distributed with zero expectation and unit variance. It is further assumed $\hat{\theta}_{ik}$ is conditionally unbiased (i.e. $E(\hat{\theta}_{ik}|\theta_{ik}) = \theta_{ik}$, which implies the measurement errors are uncorrelated with the true trait levels. It then follows the squared correlation between $\hat{\theta}_k$ and $\theta_k$ is

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

which is taken as the marginal reliability of the factor score estimates (see Ferrando and Lorenzo-Seva, in press). Denote now by $y$ the external variable or criterion also assumed to be scaled with zero mean and unit variance and by $\rho_{(\hat{\theta}_k,y)}$ the correlation between the $k_{th}$ factor score estimates and the criterion (i.e. the raw validity coefficient). From the results above it follows that the disattenuated correlation between the estimated primary factor scores and the criterion

$$\hat{\rho}_{(\theta_k,y)} = \frac{\rho_{(\hat{\theta}_k,y)}}{\sqrt{\rho_{(\hat{\theta}_k,\hat{\theta}_k)}}} \tag{3}$$

is an unbiased estimate of the corresponding correlation between the true primary scores and the criterion (i.e. the true validity coefficient). Now let $\gamma_{kg}$ be the loading of the $k$ primary factor on the general factor (i.e. the second-order loading). If the null model is correct, the following result should hold

$$\frac{\hat{\rho}(\theta_1,y)}{\gamma_{1g}} = \cdots \frac{\hat{\rho}(\theta_k,y)}{\gamma_{kg}} = \cdots \frac{\hat{\rho}(\theta_q,y)}{\gamma_{qg}}. \tag{4}$$

In words, equation 4 means the primary factors relate to the external variable in the same proportion to how they relate to the general factor. So, after correcting for this proportionality, the corrected indices should all be equal (i.e. no differential validity). To test this result, **unival** uses the following schema. First, it provides the Bootstrap-based confidence interval for each of the scaled coefficients in equation 4. Second, the median value of the scaled coefficients is obtained, and the most extreme scaled value is subtracted from the median. Next, a confidence interval for this difference is obtained via Bootstrap resampling, and a check is made to see whether the zero value falls within this interval or not. This second procedure provides a single difference statistic regardless of the number of primary factors.

If the equality test is found not tenable, then the alternative explanation (i.e. differential validity) is the unique parts of the primary factors are still differentially related to the external variable beyond the relations that are mediated by the general factor. If this were so, validity information would be lost if the unidimensional model was chosen instead of the multiple model.

We turn now to incremental validity. The starting point of the proposal by (Ferrando and Lorenzo-Seva, 2019) was based on two results. First, the score estimates on the general factor are a linear composite of the score estimates on the primary factors in which the weights aim to maximize the accuracy of the general scores. And second, the multiple-regression composite, which is also based on the primary factor score estimates, has weights aimed at maximizing the correlation with the external variable. In a truly unidimensional solution both sets of weights are expected to be proportional, and the predictive power of the general score estimates and the primary score estimates to be the same. More in detail, (Ferrando and Lorenzo-Seva, 2019) proposed correcting the primary factor score estimates for measurement error, and then obtained single and multiple corrected correlation estimates whose expected values were the same under the null model above. Under the alternative hypothesis, on the other hand, the corrected multiple correlation (denoted by $R_c$ ) was expected to be larger than the single correlation based on the general scores (denoted by $\hat{\rho}_{\hat{\theta}_g y}$). The procedure implemented in **unival** for testing the null hypothesis of no incremental validity is to compute the difference $R_c - \hat{\rho}_{\hat{\theta}_g y}$, obtain the Bootstrap confidence interval for this difference, and check whether the zero value falls within the interval or not. If the null hypothesis is rejected, the alternative explanation (i.e. incremental validity) is the primary score estimates contain additional information allowing the multiple prediction based on them to be significantly better than the prediction based only on the general scores.

## New methodological contributions

The present article extends the original proposal by (Ferrando and Lorenzo-Seva, 2019) in two directions. First, the procedure can now be correctly used with types of score estimate other than those considered initially. Second, an approximate procedure is proposed for testing essential unidimensionality against a solution in only two correlated factors.

As for the first point, the original proposal is based on factor score estimates behaving according to the assumptions derived from equation 1. Appropriate scores of this type are mainly maximum-likelihood (ML) scores, which, in the linear FA model are known as Bartlett's (1937) scores (see Ferrando and Lorenzo-Seva, in press, for a discussion). However, other types of scores are in common use in FA applications. In particular, Bayes Expected-A-Posteriori (EAP) scores have a series of practical advantages in nonlinear FA applications (Bock and Mislevy, 1982) and are, possibly, the most commonly used scoring schema for this type of solution. EAP scores, however, are always inwardly biased (i.e. regressed towards the mean) and so do not fulfill the basic assumptions on which the original procedure was based.

Simple adaptations and corrections of the existing procedures can be obtained by viewing the EAP scores as the result of shrinking the ML scores towards the zero population mean so the shrinkage factor is the marginal reliability (Bock and Mislevy, 1982). By using this concept in the assessment of differential validity, it follows that the expected value of the raw correlation between the EAP score estimates for the $k$ factor and $y$ is given by

$$E(r_{(\hat{\theta}_{kEAP},y)}) = \frac{\rho_{(\theta_k,y)}}{\sqrt{1+E(Var(\varepsilon_{ik}|\theta_{ik}))}} \tag{5}$$

Indeed, the conditional variances in the denominator of 5 are not known, because they are based on the ML unbiased estimates. However, as the number of items increases, the posterior distribution approaches normality (Chang and Stout, 1993), and the posterior standard deviation (PSD) associated with the EAP estimate becomes equivalent to an asymptotic standard error (Bock and Mislevy, 1982). So, for factors defined, say, by 8 or more items, the following correction is expected to lead to appropriate disattenuated validity coefficients

$$\hat{\rho}_{(\theta_k,y)} = r_{(\hat{\theta}_{kEAP},y)}\sqrt{1 + E(PSD^2(\theta_{ik})}. \tag{6}$$

For very short item sets, the PSDs can be noticeably smaller than the standard errors because of the additional information contributed by the prior. The strategy proposed in this case is first to approximate the amounts of information from the PSDs by using the approximate relation (Wainer and Mislevy, 2000, p. 74)

$$PSD(\hat{\theta}) \cong \frac{1}{\sqrt{I(\hat{\theta}+1)}} \tag{7}$$

and then to use the modified correction

$$\hat{\rho}_{(\theta_k,y)} = r_{(\hat{\theta}_{kEAP},y)}\sqrt{1 + E(\frac{1}{I(\hat{\theta}_{ik})}}. \tag{8}$$

Once the EAP-based disattenuated validity estimates have been obtained, they are used in the contrast 4 in the same way as those derived from the ML scores.

We turn now to incremental validity. If EAP scores are used, the corrected estimate based on the general factor score estimates (denoted by $\hat{\rho}_{\hat{\theta}_g y}$) can be obtained as

$$\hat{\rho}_{(\theta_g,y)} = r_{(\hat{\theta}_{gEAP},y)}s_{(\hat{\theta}_{gEAP})}(1 + E(\frac{1}{I(\hat{\theta}_{ik})})) \tag{9}$$

or, if the PSD approximation is used

$$\hat{\rho}_{(\theta_g,y)} = r_{(\hat{\theta}_{gEAP},y)}s_{(\hat{\theta}_{gEAP})}(1 + E(PSD^2(\theta_{ik}))) \tag{10}$$

where $s_{(\hat{\theta}_{gEAP})}$ is the standard deviation of the EAP score estimates. As for the multiple estimate based on the primary factor scores (denoted by $R_c$), only the covariances between the score estimates and the criterion must be corrected when EAP estimates are used instead of ML estimates (see Ferrando and Lorenzo-Seva, 2019). EAP-based unbiased estimates of these covariances can be obtained as

$$\hat{C}ov_{\theta_k,y} = Cov_{(\hat{\theta}_{kEAP},y)}[1 + E(PSD^2(\theta_{ik}))] \tag{11}$$

or, by using the PSD-to-Information transformation if the number of items is very small

$$\hat{C}ov_{\theta_k,y} = Cov_{(\hat{\theta}_{kEAP},y)}[1 + E(\frac{1}{I(\hat{\theta}_{ik})})]. \tag{12}$$

Once the vector with the corrected covariances has been obtained, the rest of the procedure is the same as when it is based on ML score estimates.

Overall, the basis of the proposal so far discussed is to: (a) transform the EAP scores so they (approximately) behave as ML scores; (b) transform the PSDs so they will be equivalent to standard errors, and (c) use the transformed results as input in the standard procedure. The transformations are very simple, and the proposal is expected to work well in practical applications, as the simulation study below suggests. However, unstable or biased results might be obtained if the marginal reliability estimate used to correct for shrinkage was itself unstable or biased, or if the PSDs were directly used as if they were standard errors and the contribution of the prior was substantial.

This approximate procedure is expected to be useful in practice, because in many applications decisions must be taken about using one or two common factors. The problem in this case is a second-order solution can only be identified with three or more primary factors, and so, the initial proposal cannot be used in the bidimensional case. An approximate approach, however, can be used with the same rationale as in the original procedure.

Consider two matrices of factor score estimates (either ML or EAP): an $N \times 2$ matrix containing the estimates obtained by fitting the correlated two-factor solution, and an $N \times 1$ matrix containing the score estimates obtained by fitting the unidimensional (Spearman's) model to the item scores. Next, consider the following regression schemas in which the primary factor score estimates in the $N \times 2$ matrix are corrected for measurement error. The first regression is of the unidimensional score estimates on the corrected primary factor score estimates. The second is the regression of the criterion on the same corrected factor score estimates. Now, if the unidimensional solution is essentially correct in terms of validity, then the profiles of weights for predicting the general scores and those for predicting the criterion are expected to be the same except for a proportionality constant. Denoting by $\beta_g 1$ and $\beta_g 2$ the weights for predicting the general scores from the corrected primary estimates, and by $\beta_y 1$ and $\beta_y 2$ the corresponding weights for predicting the criterion, the contrast we propose for testing the null hypothesis no differential validity is

$$\frac{\beta_g 1}{\beta_y 1} = \frac{\beta_g 2}{\beta_y 2} \tag{13}$$

and is tested by using the same procedure as in equation 4.

With regards to incremental validity, the null hypothesis of essential unidimensionality indicates both linear composites will predict the criterion equally well. So, if we denote by $y'_g$ the composite based on the $\beta_g 1$ and $\beta_g 2$ weights, and by $y'_y$ the composite based on the $\beta_y 1$ and $\beta_y 2$ weights, the test of no incremental validity is based on the contrast $r(y'_y, y) - r(y'_g, y)$, and is tested in the same way as the standard contrast above.

## The unival package details

The current version of the (**unival**) package, which is available through CRAN, contains one main function (and additional internal functions) for implementing the procedures described in the sections above. Further details on the theoretical bases of **unival** are provided in (Ferrando and Lorenzo-Seva, 2019). The function usage is as follows.

```
unival(y, FP, fg, PHI, FA_model = 'Linear', type, SEP, SEG, relip, relig,
percent = 90, display = TRUE)
```

- `y`, the related external variable,
- `FP`, the primary factor score estimates,
- `fg`, the general or second-order factor score estimates. This argument is optional except when two primary factors are specified. In this case, second-order general score estimates cannot be obtained,
- `PHI`, inter-factor correlation matrix,
- `FA_model`, Which FA-model was used for calibration and scoring. Available options are: "Linear" (by default) or "Graded". The Graded option refers to the nonlinear FA model, in which item scores are treated as ordered-categorical variables, and includes binary scores as a specific case,
- `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 estimation will be assumed,
- `SEP`, Standard Errors (ML scores) or PSDs (EAP scores) for primary factor scores (only required when the "Graded" option is used),
- `SEG`, Standard Errors (ML scores) or PSDs (EAP scores) for the general factor (only required when the "Graded" option is used),
- `relip`, the marginal reliabilities of the primary factor scors estimates. Optional when three or more primary factors are specified; otherwise, the user should provide them,
- `relig`, the marginal reliability of the general factor score estimates (optional).

The data provided should be a data frame or a numerical matrix for input vectors and matrices, and numerical values for the arguments containing a single element, like `relig`. The package imports three additional packages: **stats** (R Core Team, 2018), **optimbase** (Bihorel and Baudin, 2014) and **psych** (Revelle, 2018), for internal calculations (e.g. using the '`fa`' function from **psych** package for performing the FA calibration).

Since the function requires the factor score estimates as input, these estimates must be obtained from the raw data (i.e. the raw item scores) before **unival** is used. We recommend the non-commercial FACTOR program (Lorenzo-Seva and Ferrando, 2013) to obtain EAP estimates under the linear and the graded FA model, or the **mirt** R package (Chalmers, 2012) to obtain ML and EAP estimates for both models. FACTOR also provides PSDs for the EAP scores. Finally, both programs provide marginal reliability estimates for the chosen factor scores.

## Simulation studies

The sensitivity of the procedures proposed in **unival**, for both differential and incremental validity, depends on two main factors. The first is the relative strength of the relations between (a) the general factor scores and the external variables, and (b) the primary factor scores and the external variable. The second is the extent of the agreement between the relations between the unique parts of the primary factor and the external variables and the relations between the primary factor scores and the general factor. In summary, differential and incremental validity are expected to be clearly detected when (a) the primary factor scores are more strongly related to the external variable than to the general scores, and (b) the relation between the unique parts of the primary scores and the external variables is the opposite of the relation between the corresponding factors and the general factor. The opposite condition: (a) a general, dominant factor relates more strongly to the external variable than the primary factors do; and (b) a similar profile of relations in which the primary factors relate to the external variable in the same way as they do with the general factor, is very difficult to distinguish from the null hypothesis on which the procedures are based.

Ferrando and Lorenzo-Seva (2019) undertook a general simulation study in which the determinants above were manipulated as independent variables together with sample and model size. The study was based on the linear FA model and Bartlett's ML score estimates. In this article we replicated the study above but we discretized the continuous item responses in five response categories (i.e. a typical Likert score) and fitted the data using the non-linear FA model, thus treating the item scores are ordered-categorical variables. In addition, the factor score estimates were Bayes EAP scores. The present study, then, considers the second potential FA model that can be used in **unival**, and assesses the behavior of some of the new developments proposed in the article (the use of Bayes scores instead of ML scores). Because the design and conditions of the study were the same as those in Ferrando and Lorenzo-Seva (2019) the results are only summarized here. Details and tables of results can be obtained from the authors. The results generally agreed quite well with those obtained in the original study except for the (unavoidable) loss of power due to categorization. More in detail, in the study under the null model, neither spurious differential nor incremental validity was detected in any of the conditions.

In the studies in which the alternative model was correct, the following results were obtained. Differential validity was correctly detected except in the least favorable cells: dominant general-factor relations and profile agreement. As for incremental validity, the loss of power was more evident, and the procedure was less sensitive than in the continuous case: when the profiles of relations agreed (i.e. when the primary factors related to the external variable in the same way as they related to the general factor), **unival** failed to detect the increments in predictive power. This result, which, to a lesser extent, had already been obtained in the original study, suggests the unique relations have already been taken into account by the general factor score estimates. So, the multiple-regression linear composite, with weights very similar to those of the general factor score composite, does not substantially add to the prediction of the external variable. Overall, then, the results of the study suggest that in low-sensitivity conditions the **unival** outcome leads to the unidimensional model being chosen even when unique relations with the criterion do in fact exist. This choice, however, is probably not a practical limitation, as in these conditions the unidimensional model is more parsimonious and can explain the validity relations well. Finally, as for the differences with the previous study, the results suggest the **unival** procedures also work well with the non-linear FA model and Bayes scores. However, as expected, the categorization of the responses leads to a loss of information which, in turn, results in a loss of sensitivity and power. The most reasonable way to compensate for this loss would probably be to use a larger number of items.

## Illustration with real data

The **unival** package contains an example dataset – SAS3f – which is a matrix containing a criterion (marks on a final statistics exam), the primary factor score estimates and the general factor score estimates in a sample of 238 respondents. Both the primary and general scores were EAP estimates obtained with the FACTOR (Lorenzo-Seva and Ferrando, 2013) program.

The instrument under scrutiny is the Statistical Anxiety Scale (SAS, Vigil-Colet et al., 2008) a 24-item instrument which was initially designed to assess three related dimensions of anxiety: Examination anxiety (EA), asking for help anxiety (AHA) and interpretation anxiety (IA). Previous studies have obtained a clear solution in three highly-related factors but have also found an essentially unidimensional solution is tenable. So, the problem is to decide whether it is more appropriate to use only single-factor scores measuring an overall dimension of statistical anxiety or it is preferable (and more informative) to use the factor score estimates in each of the three dimensions.

The only remaining argument for running **unival** with minimal input requests is the inter-factor correlation matrix between the primary factors. The example should be specified as follows:

```
> PHI = cbind(c(1,0.408,0.504),c(0.408,1,0.436),c(0.504,0.436,1))
> y = SAS3f[,1]
> FP = as.matrix(SAS3f[,2:4])
> fg = SAS3f[,5]
> unival(y = y, FP = FP, fg = fg, PHI = PHI, type = 'EAP')
```

The output from the above command is:

```
Unival: Assessing essential unidimensionality using external validity information

Differential validity assessment:

0.6012 (0.4615 - 0.7311)
0.2362 (0.0280 - 0.4172)
0.3635 (0.2390 - 0.5035)

Maximum difference

0.2377 (0.0891 - 0.3587) *

Incremental validity assessment:

0.3164 (0.2328 - 0.3944)
0.4107 (0.3362 - 0.4720)

Incremental value estimate

0.0943 (0.0203 - 0.1492) **

* Some factors are more strongly or weakly related to the criterion that can be
 predicted from their relations to the general factor
** There is a significant increase in accuracy between the prediction based on the
 primary factor score estimates and that based on the general factor score  estimates.
```

Overall, the results seem to be clear. In differential validity terms, the confidence intervals for the first and second factors do not overlap, and the zero value falls outside the maximum-difference confidence interval. The interpretation is the primary factors relate to the criterion in ways that cannot be predicted from their relations with the general factor. More specifically, the first factor (AHA) seems to be more strongly related, and the second factor (IA) more weakly related to the criterion than could be predicted by their relations with the general factor.

Incremental-validity results are also clear: the prediction of the criterion based on the primary factor estimates clearly outperforms the prediction that can be made from the general factor score estimates when the regressions are corrected for measurement error. Note in particular the zero value falls well outside the confidence interval of the incremental validity estimate. To sum up, it is clear both information and predictive power will be lost in this example if the single or general factor score estimates are used as a summary of the estimates based on the three anxiety factors. So, in terms of validity, the FA solution in three correlated factors seems to be preferable.

## Concluding remarks

In the FA literature, several authors (e.g. Carmines and Zeller, 1991; Floyd and Widaman, 1995; R., 1972; Vigil-Colet et al., 1988) have pointed out the dimensionality of a set of item scores cannot be decided solely in internal terms. Rather, the ultimate criterion for judging what the most appropriate solution is should be how the scores derived from this solution relate to relevant external variables. In spite of this, however, external information is rarely used in FA-based assessments. One explanation for this state of affairs is, indeed, the difficulty of collecting additional relevant external measures. Apart from this, however, clear and rigorous procedures on how to carry out this assessment have only been proposed recently and, so far, have not been implemented in non-commercial software. For this reason, we believe **unival** is a useful additional tool for researchers who use FA in psychometric applications.

**unival** has been designed to work with scores derived from an FA solution rather than from raw item scores, and this has both shortcomings and advantages. Thus, at the minimal-input level, potential users of the program have to be able to carry out factor analyses with other programs, and, particularly, to obtain factor score estimates. Furthermore, they need to know what types of score have been computed by the program. More advanced **unival** usages require users to know how to obtain marginal reliability estimates for the factor scores or how to perform second-order factor analysis. To sum up, the program is designed for practitioners with some level of proficiency in FA. In principle, this is a potential shortcoming but does not restrict the usefulness of the program. As described above, all the input required by **unival** can be obtained from non-commercial FA packages, some of which are also quite user friendly.

The choice of the factor scores as input, on the other hand, makes the program extremely flexible and versatile. **unival** can work with scores derived from standard linear FA solutions or from non-linear solutions (which include the multidimensional versions of the graded-response and the two-parameter IRT models). Furthermore, users can choose to provide the minimal input options, or can tailor the input by choosing the type of marginal reliability estimate to be used in the error corrections or the general factor score estimates on which the analyses are based (second-order factor scores or scores derived from directly fitting the unidimensional model). No matter how complex the model or input choices are, however, the output provided by **unival** is extremely simple and clear to interpret, as the illustrative example shows.

## Acknowledgments

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

## Bibliography

M. S. Bartlett. The statistical conception of mental factors. *British Journal of Psychology*, 28:97–104, 1937. URL https://doi.org/10.1111/j.2044-8295.1937.tb00863.x. [p3]

A. Beauducel, C. Harms, and N. Hilger. Reliability estimates for three factor score estimators. *International Journal of Statistics and Probability*, 5(6):94–107, 2016. URL https://doi.org/10.5539/ijsp.v5n6p943. [p1]

S. Bihorel and M. Baudin. *Optimbase: R Port of the Scilab Optimbase Module*, 2014. URL https://CRAN.R-project.org/package=optimbase. R package version 1.0-9. [p5]

R. D. Bock and R. J. Mislevy. Adaptive eap estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4):431–444, 1982. URL https://doi.org/10.1177/014662168200600405. [p3]

E. G. Carmines and R. A. Zeller. *Reliability and Validity Assessment*, volume 17. SAGE, 1991. ISBN 9780803913714. [p7]

R. P. Chalmers. mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6):1–29, 2012. URL https://doi.org/10.18637/jss.v048.i06. [p5]

H. Chang and W. Stout. The asymptotic posterior normality of the latent trait in an irt model. *Psychometrika*, 58(1):37–52, 1993. URL https://doi.org/10.1007/BF02294469. [p3]

P. J. Curran, V. T. Cole, D. J. Bauer, W. A. Rothenberg, and A. M. Hussong. Recovering predictor–criterion relations using covariate-informed factor score estimates. *Structural Equation Modeling: A Multidisciplinary Journal*, 25(6):860–875, 2018. URL https://doi.org/10.1080/10705511.2018.1473773. [p1]

P. J. Ferrando and U. Lorenzo-Seva. Program factor at 10: Origins, development and future directions. *Psicothema*, 29:236–241, 2017. URL https://doi.org/10.7334/psicothema2016.304. [p1]

P. J. Ferrando and U. Lorenzo-Seva. Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. *Educational and Psychological Measurement*, 78(5):762–780, 2018. URL https://doi.org/10.1177/0013164417719308. [p1]

P. J. Ferrando and U. Lorenzo-Seva. An external validity approach for assessing essential unidimensionality in correlated-factor models. *Educational and Psychological Measurement*, 2019. URL https://doi.org/10.1177/0013164418824755. [p1, 2, 3, 4, 5, 6]

P. J. Ferrando and U. Lorenzo-Seva. On the added value of multiple factor score estimates in essentially unidimensional models. *Educational and Psychological Measurement*, in press. URL https://doi.org/10.1177/0013164418773851. [p1, 2, 3]

P. J. Ferrando and D. Navarro-Gonzalez. Assessing the quality and usefulness of factor-analytic applications to personality measures: A study with the statistical anxiety scale. *Personality and Individual Differences*, 123(1):81–86, 2018. URL `https://doi.org/10.1016/j.paid.2017.11.014`. [p2]

F. J. Floyd and K. F. Widaman. Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment*, 7(3):286–299, 1995. URL `https://doi.org/10.1037/1040-3590.7.3.286`. [p1, 7]

A. Furnham. The development of single trait personality theories. *Personality and Individual Differences*, 11(9):923–929, 1990. URL `https://doi.org/10.1016/0191-8869(90)90273-T`. [p1]

S. J. Haberman. When can subscores have value? *Journal of Educational and Behavioral Statistics*, 33(2): 204–229, 2008. URL `https://doi.org/10.3102/1076998607302636`. [p1]

U. Lorenzo-Seva and P. J. Ferrando. Factor 9.2: A comprehensive program for fitting exploratory and semiconfirmatory factor analysis and irt models. *Applied Psychological Measurement*, 37(6):497–498, 2013. URL `https://doi.org/10.1177/0146621613487794`. [p1, 5, 6]

G. L. R. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs*, 72(2):59, 1972. [p7]

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`. [p5]

T. Raykov and G. A. Marcoulides. On studying common factor dominance and approximate unidimensionality in multicomponent measuring instruments with discrete items. *Educational and Psychological Measurement*, 78(3):504–516, 2018. URL `https://doi.org/10.1177/0013164416678650`. [p1]

S. P. Reise, W. E. Bonifay, and M. G. Haviland. Scoring and modeling psychological measures in the presence of multidimensionality. *Journal of personality assessment*, 95(2):129–140, 2013. URL `https://doi.org/10.1080/00223891.2012.725437`. [p1]

S. P. Reise, K. F. Cook, and T. M. Moore. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In *Handbook of Item Response Theory Modeling*, pages 13–40. Routledge, 2015. [p1]

W. Revelle. *Psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois, 2018. URL `https://CRAN.R-project.org/package=psych`. R package version 1.8.10. [p5]

A. Rodriguez, S. P. Reise, and M. G. Haviland. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychological Methods*, 21(3):137–150, 2016a. URL `https://doi.org/10.1037/met0000045`. [p1]

A. Rodriguez, S. P. Reise, and M. G. Haviland. Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of personality assessment*, 98(3):223–237, 2016b. URL `https://doi.org/10.1080/00223891.2015.1089249`. [p1]

A. Vigil-Colet, U. Lorenzo-Seva, and L. Condon. Number of factors in the personality sphere: Does increase in factors increase predictability of real-life criteria? *Journal of Personality and Social Psychology*, 55(4):675–680, 1988. URL `https://doi.org/10.1037/0022-3514.55.4.675`. [p7]

A. Vigil-Colet, U. Lorenzo-Seva, and L. Condon. Development and validation of the statistical anxiety scale. *Psicothema*, 20(1):174–180, 2008. URL `https://doi.org/10.1037/t62688-000`. [p6]

H. Wainer and R. J. Mislevy. Item response theory, item calibration and proficiency estimations. In H. Wainer, editor, *Computerized Adaptive Testing: A Primer*, pages 61–101. LEA, 2000. [p4]

*Pere J. Ferrando*
*Department of Psychology*
*University Rovira i Virgili*
*Spain*
*0000-0002-3133-5466*
`perejoan.ferrando@urv.cat`

*Urbano Lorenzo-Seva*
*Department of Psychology*
*University Rovira i Virgili*
*Spain*
*0000-0001-5369-3099*
`urbano.lorenzo@urv.cat`

*David Navarro-Gonzalez*
*Department of Psychology*
*University Rovira i Virgili*
*Spain*
*0000-0002-9843-5058*
`david.navarro@urv.cat`