Personality and Individual Differences xxx (2017) xxx-xxx

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com

# Assessing the quality and usefulness of factor-analytic applications to personality measures: A study with the statistical anxiety scale

Pere J. Ferrando*, David Navarro-González

*Research Centre for Behavioural Assessment (CRAMC), 'Rovira i Virgili' University, Spain*

ARTICLE INFO

*Article history:*
Received 12 September 2017
Received in revised form 30 October 2017
Accepted 8 November 2017
Available online xxx

*Keywords:*
Personality measurement
Factor analysis
Item response theory
Factor score estimates
Reliability
Anxiety towards statistics

ABSTRACT

Factor analysis (FA) is the most widely used modeling approach for developing and assessing psychometric personality measures. Furthermore, the appropriateness of an FA application of this type is generally judged on the sole basis of model-data fit, a criterion which is clearly insufficient. This article proposes a multi-faceted approach for assessing (a) the strength and replicability of the factorial solution, (b) the accuracy and effectiveness of the factor score estimates, and (c) the closeness to unidimensionality in measures that were initially designed to be single-trait. The proposal was applied to a measure of statistical anxiety, the SAS, and the main results were the following: (a) both the unidimensional and the oblique solutions were well defined and replicable, and they led to accurate factor score estimates; and (b) unidimensional-based scores were effective over the full practical range of trait values whereas the ranges of the more specific factors in the oblique solution were narrower. It is submitted that the use of the proposal and the accompanying criteria has important advantages and can help to raise standards in FA applications in personality.



## 1. Introduction

In a general sense, factor analysis (FA) is the most widely used model for the analysis and development of personality measures. First, most traditional personality questionnaires were developed by using the standard linear FA model (e.g. Eysenck & Eysenck, 1969). Furthermore, item response theory (IRT) models that are commonly used in personality measurement, such as the one- and two-parameter models and the graded response model, can be formulated as FA models, and this formulation has important advantages especially when fitting multidimensional measures (e.g. Ferrando & Lorenzo-Seva, 2013).

In this article we shall consider full psychometric FA applications to personality measures based on a two-stage approach. In the first stage (calibration), the structure of the test is assessed and the item parameters are estimated. In the second stage (scoring) individual trait estimates (factor score estimates in the FA context) are obtained.

The appropriateness of an FA application of the type discussed so far is generally assessed by means of a goodness of fit investigation. Furthermore, because FA is a particular type of structural equation model (SEM), rigorous goodness-of-fit assessment in FA can be based on the same procedures that are used with SEMs in general (see Ferrando & Lorenzo-Seva, 2017a). In fact, goodness-of-fit assessment has become so fundamental to personality measurement that a full special issue of PAID (May 2007) was devoted to it.

Acceptable goodness of model-data fit, however, provides insufficient information for judging the quality and practical usefulness of an FA application. Good model-data fit results are perfectly compatible with weak FA structures that are very unlikely to replicate across different samples and which, in turn, yield factor score estimates that are indeterminate and unreliable, and cannot provide accurate individual measurement. To be of quality and practical usefulness, then, an FA application has to meet three standards: (a) acceptable model-data fit, (b) a clear, strong, and replicable factor structure, and (c) factor score estimates that provide accurate measurement over the range of trait levels for which the test is intended. Standards (b) and (c) are dealt with in this article.

A review of FA studies in personality measurement (e.g. Peterson, 2000; Reise, Bonifay, & Haviland, 2013) suggests that meeting the standards above is more the exception than the rule. In some cases poor starting designs might be the root of the problem. In other cases, the root might be in the use of inappropriate FA models. More specifically, the use of linear FA under conditions in which the item-factor regressions are non-linear is expected to produce artifactual 'curvature' factors that have no substantive meaning (see e.g. Ferrando & Lorenzo-Seva, 2013). Finally, over-reliance on goodness of fit criteria is another plausible reason. In order to attain an acceptable statistical fit, many measures have to include additional weak, minor, and ill-defined factors with little substantive interest (e.g. Reise et al., 2013; Reise, Cook, & Moore, 2015).

Although a variety of indices aimed at assessing standards (b) and (c) above have been proposed, coherent and organized frameworks for judging the quality and usefulness of an FA solution have only appeared recently. Rodriguez, Reise, and Haviland (2016a, 2016b) made a proposal of this type in the context of bifactor solutions,

* Corresponding author at: Universidad 'Rovira i Virgili', Facultad de Psicologia, Carretera Valls s/n, 43007 Tarragona, Spain.
*Email address:* perejoan.ferrando@urv.cat (P.J. Ferrando)

https://doi.org/10.1016/j.paid.2017.11.014
0191-8869/

whereas Ferrando and Lorenzo-Seva (2017b) and Ferrando, Navarro-González, and Lorenzo-Seva (2017) made a similar proposal in the context of the correlated-factors model. This last proposal is more general, and can be used for all sorts of FA solution, with both the linear FA model and the IRT-based FA models.

### *1.1. Objectives*

This article aims to (a) provide a non-technical and conceptual discussion of the general proposal discussed above, and (b) describe an application to a personality measure. It has a triple purpose: illustrative, substantive, and instrumental. At the illustrative level, we discuss (c) the rationale of the indices proposed, (b) how the results they provide are interpreted and, above all, (c) how the quality and practical usefulness of an FA application in personality should be assessed. In point (c) in particular, we discuss the extent to which scores based on a unidimensional FA model are interpretable and psychometri cally justified in measures considered to be multidimensional.

At the substantive level, we use the proposal to assess the properties and functioning of a popular measure of anxiety. Finally, at the instrumental level, the article provides practical information on how the proposal can be applied by using a non-commercial program.

### *1.2. Review of indices and reference values*

As stated above, the discussion provided here is only conceptual and non-technical. Technically-oriented presentations can be found in Ferrando and Lorenzo-Seva (2017b), and Ferrando et al. (2017).

#### *1.2.1. Strength and replicability of the factor solution*

Minimal rules for adequately defining a factor have been provided in the FA literature. Statistically, a factor needs a minimal of three non-zero loadings to be identified (Anderson & Rubin, 1956, p. 120). However, McDonald (1985) noted that if a factor was defined by fewer than four items with loadings above 0.30, improper solutions and Heywood cases were likely to occur. So, McDonald's recommendation seems to be a good starting rule. Beyond that, however, numerical indices are not usually considered for assessing the strength of a given FA solution.

Hancock and Mueller (2001) proposed an index, which they called *H*, for assessing the extent to which a factor is well represented by a set of items. Ferrando and Lorenzo-Seva (2017b) generalized it to the case of multiple oblique solutions, and called the general index *G-H*. Essentially, *G-H* is an estimate of the squared multiple correlation between the factor that is measured and its indicators (items), so it measures the maximal proportion of the variance of the factor that can be accounted for by the items it is measured by. More substantively, *G-H* assesses two main properties of the FA solution: (a) the quality of the items as indicators of the factor, and (b) the expected replicability of the solution across studies. So, low *G-H* values are indicators of a weak, ill-defined solution that is unlikely to replicate across different samples or studies. As for reference values, Hancock and Mueller proposed 0.70 as a minimal value if the factor is to be regarded as well represented, whereas Rodriguez et al. (2016b) and Ferrando and Lorenzo-Seva (2017b) raised this to 0.80, which is the minimal cut-off proposed here.

#### *1.2.2. Quality and effectiveness of the factor score estimates*

The effectiveness of the factor score estimates is a multifaceted concept which comprehends several properties (Ferrando et al., 2017). The first is the precision with which the latent trait levels can be estimated. The second is the sensitivity of the factor score estimates for differentiating individuals with different trait levels. The third is the range of trait levels at which the factor score estimates are precise and provide good precision and differentiation.

The standard index of score effectiveness is the coefficient of marginal reliability which is both a measure of precision and a measure of sensitivity (see Ferrando et al., 2017).It also indicates the degree of relation between the factor score estimates and the latent levels in the factor they estimate. So, high reliability values mean that respondents can be accurately measured and effectively differentiated on the basis of their score estimates, and that the factor score estimates are good proxies for the corresponding latent factor.

The coefficient of marginal reliability can be viewed as: (a) the ratio of variance of the latent factor or trait levels over the variance of the estimated factor scores, and (b) the squared correlation between the latent factor or trait levels and the estimated factor scores (e.g. Brown & Croudace, 2015, Ferrando and Lorenzo-Seva, 2017a, b). These are two standard definitions of a reliability coefficient in general, and, for this reason, we consider that the same reference values that are used for any standard reliability coefficient can also be used for marginal reliability. A minimal value of 0.80 seems to be a reasonable cut-off if the factor score estimates are to be used for individual measurement (Ferrando and Lorenzo-Seva, 2017a, b).

In nonlinear (IRT) FA models, the reliability varies depending on the level of the respondent, so in these models, the marginal reliability above is an average of the individual or conditional reliabilities (see Ferrando, 2003 Ferrando et al., 2017). If the individual reliabilities remain relatively uniform across the different levels, the marginal reliability is representative of the overall precision of the scores (e.g. Brown & Croudace, 2015) and the test is considered to measure about equally well at all levels. However, this is not so in general (Ferrando, 2003).

Most personality tests are designed to be broad bandwidth measures and aim to accurately measure most individuals from the population for which the test is intended (Ferrando, 2003). What we propose for assessing if this is so is a graphical approach that estimates the interval of trait levels at which the factor score estimates are effective. Consider the graphic display of the conditional reliabilities against the factor score estimates, and define a minimally acceptable cut-off value of, say, 0.80. This cut-off is a horizontal line parallel to the trait axis, and the range of effectiveness can be defined as the trait interval at which the reliabilities are above this line. The usefulness of this proposal is discussed in detail in the empirical study.

A final auxiliary index we would like to consider is the so called "expected percentage of true differences" (EPTD; Ferrando et al., 2017), which reflects the percentage of observed differences between the factor score estimates that are in the same direction as the corresponding latent differences. So EPTD addresses a somewhat different aspect of effectiveness: it is not about the size of the differences that can be detected (i.e. reliability) but about the proportion of differences (of any size) that are in the correct direction. The higher this proportion, the better individuals can be consistently differentiated or ordered along the factor continuum on the basis of their factor score estimates. Values of EPTD above 0.90 seem a minimal requirement if the factor score estimates are to be used for individual assessment.

#### *1.2.3. Closeness to unidimensionality*

Although many personality measures were initially intended and designed to be single-trait or unidimensional, subsequent FAs nearly always arrive at multidimensional solutions (Furnham, 1990; Reise et al., 2013; Reise et al., 2015), especially in those measures aimed at assessing broad-bandwidth traits. In some cases, the multiple solutions are meaningful and reach the quality standards discussed above.

In many others, however, they are the result of inappropriate FA models, (i.e. spurious evidence of multidimensionality because the linear model was used in data that required the use of a nonlinear model), or were obtained solely with the aim of achieving acceptable levels of model-data fit. In these last two cases, the resulting structures and derived scores are generally unacceptably poor, are non-interpretable or have no substantive interest. Given this scenario, it is of great interest to assess the extent to which a measure that was ini tially designed as single-trait in fact behaves as essentially unidimensional.

One of the most usual auxiliary indices for assessing closeness to unidimensionality is the proportion of total variance explained by the first principal factor (e.g. Kim and Mueller, 1978), which, according to Kim and Mueller is a criterion of substantive importance. In the case of test items that have generally large amounts of measurement error, however, this index must be highly misleading. In effect, in a typical personality test, a perfectly unidimensional solution is compatible with a modest amount of total variance explained. What is of interest is the explained common variance not the total variance.

The index we propose here is simple, informative and has been proposed in slightly different variants. The one chosen here is that by ten Ten Berge and Kiers (1991) based on minimum rank factor analysis (MRFA). The explained common variance (ECV) index is defined as the proportion of common variance explained by the first principal factor with respect to the common variance contained in the test items as estimated by MRFA. As for reference values, cut-off values between 0.70 and 0.85 have been proposed to conclude that a solution is essentially unidimensional (Ferrando and Lorenzo-Seva, 2017a, b, Rodriguez et al., 2016a, 2016b). When these cut-offs are used, many oblique solutions reported in personality are found to be compatible with an essentially unidimensional solution (Reise et al., 2013; Reise et al., 2015).

### *1.3. Current empirical study*

The proposal discussed in the section above was applied in a substantive study based on a measure of anxiety towards statistics: the Statistical Anxiety Scale (SAS). The SAS was designed to assess three related dimensions of anxiety, and also to be used as a general statistical-related anxiety measure (Vigil-Colet, Lorenzo-Seva, & Condon, 2008). So, the present proposal is particularly relevant for the study, and can provide information about three key points. First, whether the structure in three factors and the resulting score estimates attain the standards of quality proposed here. Second, whether the SAS can be used as an essentially unidimensional measure. And, finally, whether it is more appropriate to use it as a tridimensional measure or as a unidimensional measure.

## 2. Method

### *2.1. Participants*

The present sample was made up of 384 undergraduate students enrolled on a statistics course in a faculty of Psychology in Spain. There were 327 women and 56 men, between 18 and 35 years old (mean = 20.47; standard deviation = 4.89). The administration was in classroom groups and under standard instructions.

### *2.2. Measures*

The SAS is a 24-item measure intended to assess the anxiety levels of students taking a statistics course. It assesses three related components of anxiety: Examination Anxiety (8 items), Asking for Help Anxiety (8 items) and Interpretation Anxiety (8 items). All of the items are positively worded and use a five-point Likert response format, ranging from "no anxiety" (1) to "considerable anxiety" (5).

FA-based studies on the SAS generally obtained clear structures with 3 highly correlated factors (Chiesi, Primi, & Carmona, 2011; Oliver, Sancho, Galiana, and Cebrià i Iranzo, M. A., 2014; Chew & Dillon, 2014). Mainly for this last reason, the authors suggested that the SAS could also be considered an essentially unidimensional measure and be used as an overall 24-item scale (Vigil-Colet et al., 2008).

### *2.3. Analyses*

The SAS scores were fitted using the two-stage procedure discussed above. The item parameter estimates in the calibration stage were obtained by using robust unweighted least squares estimation as implemented in the FACTOR program (Lorenzo-Seva & Ferrando, 2013). The item responses were treated as ordered-categorical variables, so the CVM-FA based on the polychoric inter-item correlations was the chosen model. This model is an alternative parameterization of the multidimensional IRT graded response model and its use prevents the problems caused by applying the linear model discussed above (see Ferrando & Lorenzo-Seva, 2013). The following indices were used to assess goodness of model-data fit: Root Mean Square Error of Approximation (RMSEA); Comparative Fit Index (CFI; an incremental index), and Root Mean Square of the standardized Residuals (z-RMSR; an absolute index). 95% confidence intervals were also reported for all of these indices. Finally, when computing the tridimensional model, Promin (Lorenzo-Seva, 1999) was used as a rotation method, because the factors were expected to be correlated.

In the scoring stage, Bayes EAP factor score estimates and the corresponding standard errors were obtained for each of the solutions described above. The prior distributions for the specified factors were assumed to be standard normal. So, the effective range of both trait levels or factor scores (denoted by $\Theta$) and factor score estimates (denoted by $\hat{\theta}$) was − 3 to + 3. Finally the indices proposed here were obtained with FACTOR based on the calibration and scoring results.

## 3. Results

### *3.1. Item calibration*

Goodness of model-data fit results are in Table 1 and are quite clear. The fit of the unidimensional model is poor, especially in terms of the magnitude of residuals, whereas the fit of the tridimensional model is excellent by all the standards.

The rotated pattern for both models is in Table 2. As expected, the pattern corresponding to the 3-factor solution agrees quite well with the prescribed 'a priori' structure, with all the salient loadings (boldfaced) located in the corresponding factor.

For all the factors in Table 2, the minimal rule for adequate factor identification is clearly fulfilled. Note also that the single-factor solution exhibits positive manifold and that all its loadings are above

**Table 1**
Goodness of fit indices and confidence intervals.

| Index | Unidimensional model | Tridimensional model |
|---|---|---|
| | Value (95% confidence interval) | Value (95% confidence interval) |
| RMSEA | 0.128 (0.110–0.139) | 0.033 (0.026–0.033) |
| CFI | 0.917 (0.892–0.943) | 0.995 (0.995–0.997) |
| z-RMSR | 0.138 (0.121–0.149) | 0.042 (0.040–0.042) |

**Table 2**
Pattern Loading Matrices and inter-factor correlation matrix.

a) Pattern loadings for both models

| | Unidimensional model | Tridimensional model | | |
|---|---|---|---|---|
| Item | Global statistical anxiety | Examination anxiety | Asking for help anxiety | Interpretation anxiety |
| 1 | 0.688 | **0.649** | 0.081 | 0.089 |
| 4 | 0.713 | **0.725** | 0.029 | 0.101 |
| 9 | 0.738 | **0.803** | 0.128 | − 0.078 |
| 11 | 0.679 | **0.783** | 0.106 | − 0.110 |
| 13 | 0.633 | **0.806** | − 0.116 | 0.078 |
| 14 | 0.707 | **0.653** | 0.151 | 0.021 |
| 15 | 0.636 | **0.974** | − 0.193 | − 0.013 |
| 20 | 0.625 | **0.886** | − 0.163 | 0.032 |
| 3 | 0.702 | 0.023 | **0.907** | − 0.150 |
| 5 | 0.655 | − 0.005 | **0.598** | 0.185 |
| 7 | 0.780 | 0.007 | **0.885** | 0.003 |
| 12 | 0.772 | − 0.01 | **0.971** | − 0.087 |
| 17 | 0.784 | − 0.056 | **0.976** | − 0.022 |
| 21 | 0.764 | − 0.049 | **0.718** | 0.248 |
| 23 | 0.711 | − 0.018 | **0.949** | − 0.138 |
| 24 | 0.718 | − 0.035 | **0.727** | 0.154 |
| 2 | 0.396 | − 0.005 | − 0.174 | **0.806** |
| 6 | 0.481 | − 0.146 | 0.048 | **0.823** |
| 8 | 0.652 | 0.256 | 0.215 | **0.340** |
| 10 | 0.386 | − 0.066 | − 0.085 | **0.750** |
| 16 | 0.515 | 0.262 | 0.100 | **0.284** |
| 18 | 0.472 | − 0.013 | 0.081 | **0.592** |
| 19 | 0.632 | 0.204 | 0.187 | **0.413** |
| 22 | 0.559 | − 0.054 | − 0.037 | **0.941** |

b) Inter-factor correlation matrix

| Factors | Examination anxiety | Asking for help anxiety |
|---|---|---|
| Asking for help anxiety | 0.581 | – |
| Interpretation anxiety | 0.487 | 0.457 |

0.30. Finally, the three factors in the oblique solution are positively and substantially correlated with one another which is a necessary (but not sufficient) requirement for considering that there is a general factor underlying all the 24 SAS item responses.

We turn now to the first group of indices proposed here. The *G-H* value for the unidimensional model was 0.952, whereas for the tridimensional model the values were 0.942 (F1), 0.968 (F2) and 0.918 (F3). So, in all cases, the factors can be considered to be strong and well defined, and therefore, it is predicted that they will be replicable.

### *3.2. Individual scoring*

The results obtained at the scoring stage are in Table 3. Overall, both models show good performance in the two indices proposed here, with marginal reliability values above 0.9 and EPTD values above 90%. The first result means that the factor score estimates (a) are highly correlated with the latent factors they represent; (b) are accurate, and (c) allow the individuals of different trait levels to be effectively differentiated. The second result means that more than 90% of the differences among the factor score estimates reflect latent differences that are in the same direction, so individuals can be consistently ordered on the basis of their estimated scores. Note that marginal reliability is highest for the scores based on the unidimensional solution, but that the differences with the first two factors are small. The third factor (Interpretation anxiety) is the least reliable, which agrees with its lower *G-H* value in the previous section.

**Table 3**
Marginal reliability and expected percentage of true differences for both models.

| | Unidimensional model | Tridimensional model | | |
|---|---|---|---|---|
| Index | Global statistical anxiety | Examination anxiety | Asking for help anxiety | Interpretation anxiety |
| Marginal reliability | 0.9730 | 0.9646 | 0.9547 | 0.9096 |
| Expected percentage of true differences (EPTD) | 97.28% | 97.58% | 96.58% | 93.26% |

Fig. 1 shows the graphical range assessment for each factor, with a cut-point of 0.80 for the conditional reliability (displayed as a horizontal line in the figure). The unidimensional model seems to show good reliability in almost all the effective trait range (i.e. from − 3 to + 3) and a clear decrease only for the score estimates below $\hat{\theta} = -2.6$ .

In contrast, the tridimensional model presents three different range patterns. The "Asking for help Anxiety" scores show good precision at low trait levels (between $\hat{\theta}_1 = -2.6$ and $\hat{\theta}_1 = 0.8$); "Examination Anxiety" scores are more effective at intermediate-high levels (between $\hat{\theta}_2 = -1.0$ and $\hat{\theta}_2 = 2.8$) and finally the "Interpretation Anxiety" scores, which correspond to the 'weakest' factor, have good reliability only for trait levels above $\hat{\theta}_3 = 0.4$.

Reise and Waller (2009) noted that many clinical instruments provide accurate measurement only over a limited range of trait values, generally at the more severe end of the trait, and they coined the term "quasi-traits" to refer to this behaviour. A quasi-trait is, then, a trait that is only relevant in one direction, so scores at one end of the scale (usually the lower end) are less informative. The quasi-trait hypothesis provides a plausible explanation for the behaviour of the "Examination Anxiety" and "Interpretation Anxiety" scores. However, the impact of end (floor and ceiling) effects cannot be discarded either as a potential cause for the results above: The scores on the "Examination Anxiety" items were generally negatively skewed, whereas items scores on the other two factors were generally positively skewed.

### *3.3. Closeness to unidimensionality*

The estimated ECV value was 0.80, which means that 80% of the common variance in the SAS items can be explained by a single general statistical anxiety factor. This value is between the two cut-off values proposed in the literature, so it provides support for using the test as a total scale, as was initially proposed.

## 4. Discussion and conclusions

The results above allow the three key points raised at the beginning to be addressed, and, in the case of points 1 and 2, the answers seem to be quite clear. With regards to the first point, the three-factor structure and the resulting score estimates attain the standards of quality we proposed. As for the second, the positive-manifold structure, the substantial inter-factor correlations, the *G-H* index, and the *ECV* all suggest that the SAS can be used as an essentially unidimensional measure, despite the fact that the unidimensional model does not arrive at an acceptable fit in purely statistical terms. More in detail, the results suggest that scores derived from the unidimensional solution are psychometrically justified and interpretable as indicators of a general statistical anxiety factor.

Because the two solutions considered in the study were found to be acceptable in terms of replicability and accuracy, addressing the third point is more complex, and many issues should be considered (Reise et al., 2013). First, scores based on the oblique solution are expected to provide additional meaningful information about the sources and specific forms of statistical anxiety, and they could even

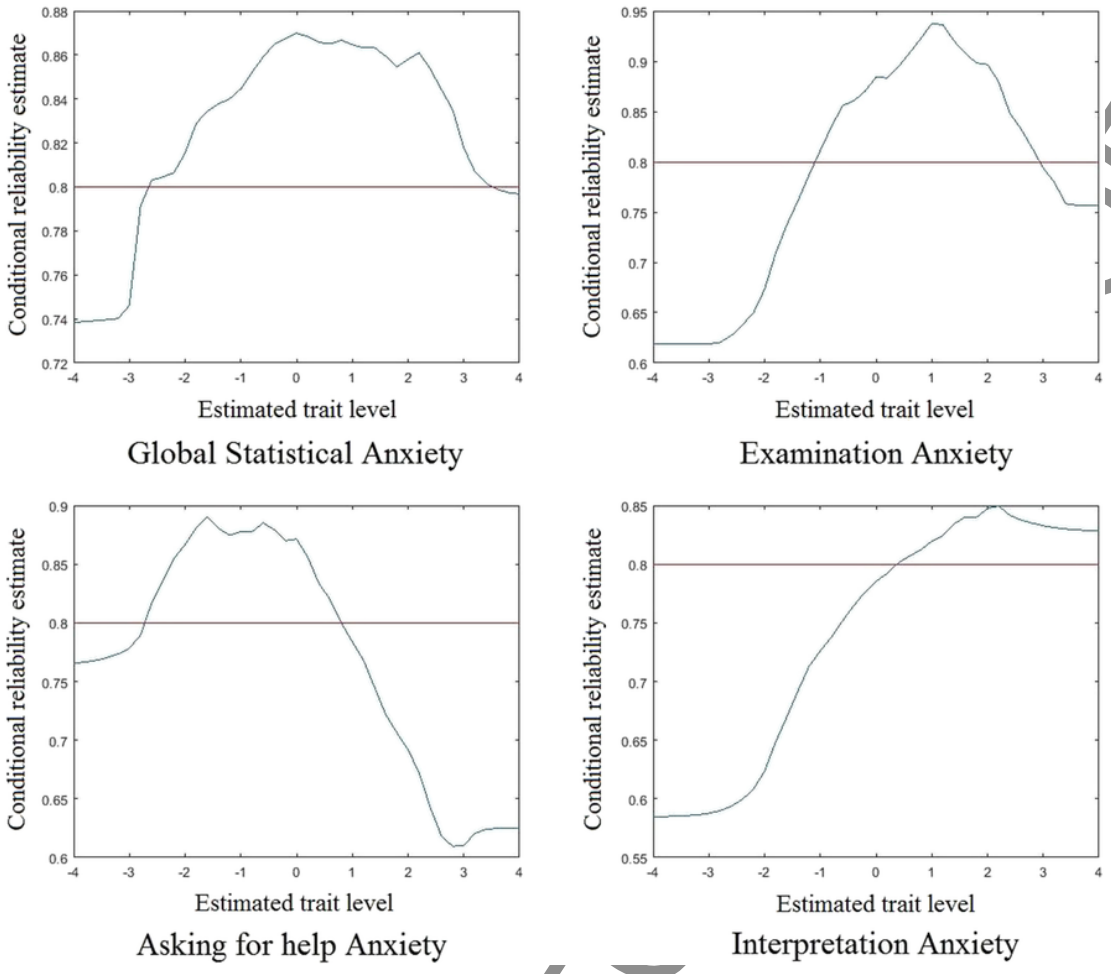


**Fig. 1.** Graphical range assessment for both models.

have distinct implications regarding individual counselling or intervention. Second, they might relate differently to relevant external criteria, which means that a validity study would be a valuable source of information for deciding how the SAS response should be scored. Third, scores based on the unidimensional solution are simpler, with slightly higher marginal reliability, and, above all, they provide accurate measurement throughout the effective range of trait values. In view of these results, the scoring based on the tridimensional model is possibly the best option in clinical assessment settings, whereas the unidimensional scoring is possibly the best option for screening purposes in the general population or when the aim is to rank individuals according to their anxiety levels. Validity studies would be needed to decide the most appropriate scoring for predictive purposes.

Finally, we turn to a more general discussion. This article is, mainly, a proposal and a substantive application. However, it is also closely related to some highly topical problems in personality measurement. First, there seems to be a growing awareness in the field that the use of excessively restricted FA models and overreliance on purely statistical model-data fit is not the way to go, and can even lead to poorer measurement outcomes (Ferrando & Lorenzo-Seva, 2017a; Rodriguez et al., 2016a, 2016b). Second, many published articles in personality deal with the dimensionality and structure of new or already existing measures (e.g. Furnham, 1990; Reise et al., 2013). However, an informal literature review clearly suggests that there is a wide gap between the proposed structures and the scoring schemas that can be derived from them (which, in most cases, are not even considered). We suspect that, in many cases, the derived scores would not attain the minimum quality requirements proposed here. So, the present proposal can be considered both as an alternative schema for assessing the quality and practical interest of an FA based solution and a potential way of raising the standards of many applications.

Experience suggests that proposals such as the present one only have the chance to be used in practice if they are available in user-friendly, relatively well known and (if possible) free programs. For this reason all the proposed procedures have been implemented in the FACTOR program, and results for any application can be obtained as they were in the SAS study. FACTOR is a freeware program available at http://psico.fcep.urv.cat/utilitats/factor/.

## References

Anderson, T.W., Rubin, H., 1956. Statistical inference in factor analysis. In: Neyman, J. (Ed.), Proceedings of the third Berkeley symposium on mathematical statistics and probability. Vol. 5, University of California Press, Berkeley, CA, pp. 111–150.

Brown, A., Croudace, T., 2015. Scoring and estimating score precision using multidimensional IRT. In: Reise, S.P., Revicki, D.A. (Eds.), Handbook of item response theory modeling: Applications to typical performance assessment. Routledge, New York, pp. 307–333.

Chew, P.K.H., Dillon, D.B., 2014. Reliability and validity of the statistical anxiety scale among students in Singapore and Australia. Journal of Tropical Psychology 4.

Chiesi, F., Primi, C., Carmona, J., 2011. Measuring statistics anxiety: Cross-country validity of the statistical anxiety scale (SAS). Journal of Psychoeducational Assessment 29 (6), 559–569.
Eysenck, H.J., Eysenck, S.B.G., 1969. Personality structure and measurement. Routledge & K. Paul, London.
Ferrando, P.J., 2003. The accuracy of the E, N and P trait estimates: An empirical study using the EPQ-R. Personality and Individual Differences 34, 665–679.
Ferrando, P.J., Lorenzo-Seva, U., 2013. Unrestricted item factor analysis and some relations with item response theory. In: Technical report. Department of Psychology, Universitat Rovira i Virgili, Tarragona 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.
Ferrando, P.J., Lorenzo-Seva, U., 2017. Assessing the quality and appropriateness of factor solutions and factor score estimates in exploratory item factor analysis. Educational and Psychological Measurement(0013164417719308, in press, available online).
Ferrando, P.J., Navarro-González, D., Lorenzo-Seva, U., 2017. Assessing the quality and effectiveness of the factor score estimates in psychometric factor-analytic applications. Methodology(submitted).
Furnham, A., 1990. The development of single trait personality theories. Personality and Individual Differences 11, 923–929.
Hancock, G.R., Mueller, R.O., 2001. Rethinking construct reliability within latent variable systems. In: Cudek, R., duToit, S.H.C., Sorbom, D.F. (Eds.), Structural equation modeling: Present and future. Scientific Software, Lincolnwood, pp. 195–216.
Kim, J.O., Mueller, C.W., 1978. Factor analysis: Statistical methods and practical issues. Vol. 14, Sage.
Lorenzo-Seva, U., 1999. Promin: A method for oblique factor rotation. Multivariate Behavioral Research 34, 347–356.
Lorenzo-Seva, U., Ferrando, P.J., 2013. FACTOR 9.2: A comprehensive program for fitting exploratory and Semiconfirmatory factor analysis and IRT models. Applied Psychological Measurement 37, 497–498.
McDonald, R.P., 1985. Factor analysis and related methods. LEA, Hillsdale.
Oliver, A., Sancho, P., Galiana, L., Cebrià i Iranzo, M. A., 2014. Nueva evidencia sobre la Statistical Anxiety Scale (SAS). Anales de psicología 30, 150–156.
Peterson, R.A., 2000. A meta-analysis of variance accounted for and factor loadings in exploratory factor analysis. Marketing Letters 11, 261–275.
Reise, S.P., Bonifay, W.E., Haviland, M.G., 2013. Scoring and modeling psychological measures in the presence of multidimensionality. Journal of Personality Assessment 95, 129–140.
Reise, S.R., Cook, K.F., Moore, T.M., 2015. Evaluating the impact of multidimensionality on unidimensional item response theory model parameters. In: Reise, S.P., Revicki, D.A. (Eds.), Handbook of item response theory modeling: Applications to typical performance assessment. Routledge/Taylor & Francis Group, New York, pp. 13–40.
Reise, S.P., Waller, N.G., 2009. Item response theory and clinical measurement. Annual Review of Clinical Psychology 5, 27–48.
Rodriguez, A., Reise, S.P., Haviland, M.G., 2016. Evaluating bifactor models: Calculating and interpreting statistical indices. Psychological Methods 21, 137.
Rodriguez, A., Reise, S.P., Haviland, M.G., 2016. Applying bifactor statistical indices in the evaluation of psychological measures. Journal of Personality Assessment 98, 223–237.
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.
Vigil-Colet, A., Lorenzo-Seva, L., Condon, L., 2008. Development and validation of the statistical anxiety scale. Psicothema 20, 174–180.