Personality and Individual Differences 123 (2018) 81–86

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

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

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

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

https://doi.org/10.1016/j.paid.2017.11.014
Received 12 September 2017; Received in revised form 30 October 2017; Accepted 8 November 2017

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, 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 psychometrically 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