Entropy Fit Indices: New Fit Measures for Assessing the Structure and Dimensionality of Multiple Latent Variables

Hudson Golino\textsuperscript{1}, Robert Moulder\textsuperscript{1}, Dingjing Shi\textsuperscript{1}, Alexander P. Christensen\textsuperscript{2}, Luis Eduardo Garrido\textsuperscript{3}, Maria Dolores Nieto\textsuperscript{4}, John Nesselroade\textsuperscript{1}, Ritu Sadana\textsuperscript{5}, Jotheeswaran Amuthavalli Thiyagarajan\textsuperscript{5}, & Steven M. Boker\textsuperscript{1}

\textsuperscript{1}Department of Psychology, University of Virginia; \textsuperscript{2}Department of Psychology, University of North Carolina at Greensboro; \textsuperscript{3}Department of Psychology, Pontificia Universidad Catolica Madre y Maestra; \textsuperscript{4}Department of Psychology, Universidad Autónoma de Madrid; \textsuperscript{5}Department of Ageing and Life Course, World Health Organization

\textbf{ABSTRACT}

The accurate identification of the content and number of latent factors underlying multivariate data is an important endeavor in many areas of Psychology and related fields. Recently, a new dimensionality assessment technique based on network psychometrics was proposed (Exploratory Graph Analysis, EGA), but a measure to check the fit of the dimensionality structure to the data estimated via EGA is still lacking. Although traditional factor-analytic fit measures are widespread, recent research has identified limitations for their effectiveness in categorical variables. Here, we propose three new fit measures (termed entropy fit indices) that combines information theory, quantum information theory and structural analysis: Entropy Fit Index (EFI), EFI with Von Neumman Entropy (EFI.vn) and Total EFI.vn (TEFI.vn). The first can be estimated in complete datasets using Shannon entropy, while EFI.vn and TEFI.vn can be estimated in correlation matrices using quantum information metrics. We show, through several simulations, that TEFI.vn, EFI.vn and EFI are as accurate or more accurate than traditional fit measures when identifying the number of simulated latent factors. However, in conditions where more factors are extracted than the number of factors simulated, only TEFI.vn presents a very high accuracy. In addition, we provide an applied example that demonstrates how the new fit measures can be used with a real-world dataset, using exploratory graph analysis.

\textbf{KEYWORDS}

Fit measures; Dimensionality Analysis; Information Theory; Network Psychometrics; Exploratory Graph Analysis
Introduction

Estimating the dimensionality or the number of underlying dimensions (or factors) in multivariate datasets is an important step in psychology and related areas. In the health and social sciences areas, assessing the dimensionality of scales, tests, and questionnaires is one of the first steps on the path to understanding multivariate data. Parallel, exploratory factor, and principal component analysis (e.g., eigenvalues and scree plots) are the most common techniques implemented in commercial software and among the most common strategies used to identify the dimensions of the data across research areas. Despite their long history and widespread application, simulation studies have consistently shown that traditional factor-analytic techniques present important limitations (Crawford et al., 2010; Garrido, Abad, & Ponsoda, 2013; Hayton, Allen, & Scarpello, 2004; Keith, Caemmerer, & Reynolds, 2016; Ruscio & Roche, 2012).

Recently, a new dimensionality assessment technique termed Exploratory Graph Analysis (EGA: Golino & Epskamp, 2017) was proposed. EGA has shown a number of advantages when compared to traditional factor-analytic techniques (Golino & Demetriou, 2017; Golino & Epskamp, 2017). The technique estimates the number of dimensions by first modeling multivariate data as a network of partial correlations (using the Gaussian graphical model, GGM; Lauritzen, 1996) or zero-order correlations (using the triangulated maximally filtered graph approach, TMFG; Massara, Di Matteo, & Aste, 2016). Then, the technique applies a community detection algorithm for weighted networks (i.e., Walktrap; Pons & Latapy, 2006), which identifies the number of dimensions in the network using a random walk process. Not only does EGA present higher accuracy than many traditional methods (i.e., percentage of correct estimates across different simulated conditions), but it also produces a visual guide (network plot) that indicates the number of dimensions to retain, showing which items cluster together and their level of association, all without the need of any specification from researchers (Golino et al., in press). Although useful, EGA lacks a fit index to compare different dimensionality structures between EGA methods (GGM and TMFG), EGA and factor-analytic methods, and EGA and theoretical structures. Moreover, community detection algorithms often have parameters (e.g., steps in the Walktrap algorithm) that adjusts the dimensions the algorithm will find. Therefore, a fit measure that can be used to automatically adjust these parameters for optimal performance could greatly improve EGA’s precision.

Applied researchers and methodologists use fit indices to assess the factor solutions of their data (or instruments), mainly because they provide useful diagnostic information, but also because they enable the comparison of different dimensionality structures (i.e., number of factors and distribution of items per factors). Despite the usefulness of fit indices used in factor analysis to estimate the number of underlying dimensions, there are two main limitations: First, there are only a few simulation studies investigating the accuracy of the traditional fit indices (such as Comparative Fit Index and Root Mean Square Error of Approximation or CFI and RMSEA, respectively) to estimate the correct number of factors (Barendse, Oort, & Timmerman, 2015; Clark & Bowles, 2018; Frazier & Youngstrom, 2007; Garrido, Abad, & Ponsoda, 2016; Yang & Xia, 2015). Second, they produce poor dimensionality estimates for skewed categorical variables (Clark & Bowles, 2018; Garrido et al., 2016). This makes there a need for fit indices that mitigate these limitations.

The goals of the paper are twofold. First, we introduce a new family of fit indices
termed Entropy Fit Indices (EFIs) will be presented: an index that is computed using the raw data (Entropy Fit Index; EFI), and two indices that is computed using the correlation matrix (Entropy Fit Index with Von Neumann Entropy and Total Entropy Fit Index; vnEFI and TEFI, respectively). These new fit indices can be used to assess the degree of disorder (or uncertainty, using information theory terms) of a set of random variables for structures with different skew of variables, number of factors, different distribution of items in factors, and placement of items in each factor. Lower values of the EFIs indicate lower disorder (or uncertainty) of a system of variables, indicating a higher probability that a given structure represents the best organization of these variables. Second, we use a Monte Carlo simulation method to compare the EFIs to some of the most commonly used fit indices in factor analysis: CFI, RMSEA, Tucker-Lewis Index (TLI), and Standardized Root Mean Square Residual (SRMR).

This article is organized as follows: The first section briefly introduces several traditional factor-analytic fit indices and their usefulness in estimating the number of underlying dimensions in multivariate data. The second section briefly reviews EGA, and summarizes the main findings of previous simulation studies. The third section briefly introduces the concept of entropy, its properties and estimation methods, and the new EFIs. The fourth section presents a Monte Carlo simulation, where the number of factors, variables per factor, interfactor correlations, factor loadings and skewness of the variables are systematically controlled. The fit indices will be computed when a given structure reflects: (1) exactly the simulated factor structure, (2) the correct number of factors, but with a random placement of items per factor, and (3) an incorrect number of factors (underfactoring or overfactoring). The fifth section uses an empirical example, from a large scale study on aging, to demonstrate the use of EFI. Finally, the last section presents a discussion of the results, limitations of the current study, and future directions.

Fit Indices used in Dimensionality Assessment

Fit indices derived from factor analyses models have also been proposed as tools to estimate latent dimensions in multivariate data (Santos, Vagos, & Rijo, 2018; Schermelleh-Engel, Moosbrugger, & Müller, 2003; Ventimiglia & MacDonald, 2012). These fit indices include, among others, the CFI (Bentler, 1990), TLI (Tucker & Lewis, 1973), RMSEA (Steiger & Lind, 1980), and SRMR (Joreskog & Sorbom, 1981). The first three indices are based on the noncentrality parameter (λ_M) of the specified model, which is computed as $\chi^2_M - df_M$, where $\chi^2_M$ is the chi-square statistic that tests the equivalence of the population covariance matrix of observed variables and the model-implied covariance matrix, and $df_M$ is the degrees of freedom of the specified model. The last fit index, SRMR, is based on the average difference between observed and model-implied covariance matrices. The CFI and TLI indices are considered incremental fit indices because they evaluate the degree to which the specified model is superior to a baseline model—usually a null model where the observed variables are proposed to be uncorrelated—in reproducing the observed covariances. In contrast, the RMSEA and SRMR are considered absolute fit indices because they offer a measure of overall fit that only takes into account the fit of the specified model. A key advantage of the RMSEA index is that, because it follows a known $\chi^2$ distribution asymptotically, confidence intervals can be estimated around the point estimate, allowing for formal hypothesis tests to be conducted by jointly considering the point estimate and its
associated confidence interval (Chen, Curran, Bollen, Kirby, & Paxton, 2008). The formulas for the aforementioned fit indices are shown below in Equations 1 to 4:

\[ CFI = 1 - \frac{\text{max}(\lambda_M, 0)}{\text{max}(\lambda_N, \lambda_M)}, \]  
(1)

\[ TLI = 1 - \left( \frac{\lambda_M}{\lambda_N} \right) \left( \frac{df_N}{df_M} \right), \]  
(2)

\[ \text{RMSEA} = \text{max} \left( \sqrt{\frac{\lambda_M}{df_M(N-1)}}, 0 \right), \]  
(3)

\[ SRMR = \sqrt{\frac{\sum_{i=1}^{p} \sum_{j=1}^{i} \left( \frac{s_{ij}}{\sqrt{s_{ii}} \sqrt{s_{jj}}} - \frac{\hat{\sigma}_{ij}}{\sqrt{\hat{\sigma}_{ii}} \sqrt{\hat{\sigma}_{jj}}} \right)^2}{\frac{p(p+1)}{2}}}, \]  
(4)

where \( \lambda_N \) and \( df_N \) are the noncentrality parameter and degrees of freedom of the baseline model, \( N \) is the sample size, \( s_{ij} \) is the observed covariance, \( \hat{\sigma}_{ij} \) is the model-implied covariance, \( s_{ii} \) and \( s_{jj} \) are the observed standard deviations, \( \hat{\sigma}_{ii} \) and \( \hat{\sigma}_{jj} \) are the model-implied standard deviations, and \( p \) is the number of observed variables.

When using fit indices to estimate latent dimensionality under the framework of unrestricted (exploratory) factor models, the common rationale is that factors should be sequentially added to a model until an acceptable level of fit has been achieved. First, a one-factor model would be fitted to the data and its level of fit would be compared against pre-specified cut-off values for the fit indices. If the fit was deemed acceptable, the one-factor model would be retained; if not, an unrestricted two-factor model would be estimated and the process would be repeated, until an acceptable level of fit was achieved. Applied researchers that have employed this approach (e.g., Campbell-Sills, Liverant, & Brown, 2004; Sanne, Torsheim, Heiervang, & Stormark, 2009) have generally relied on the typical cut-off values that have been proposed for these indices in the literature, such as .90 or .95 for CFI and TLI, and .05 or .08 for RMSEA and SRMR (good and acceptable fit, respectively; Browne & Cudeck, 1992; Chen et al., 2008; Hu & Bentler, 1999).

Despite the appeal of using fit indices to estimate latent dimensionality, recent simulation studies have shown that they are not the best tools for this endeavor. Garrido et al. (2016) examined the performance of these four fit indices across a large number of conditions of continuous and categorical variables and using a wide range of cut-off values for each index. They found that the CFI and TLI indices were the most accurate, followed at a step below by RMSEA, and finally by SRMR, which produced very poor estimates across all the cut-off values in its relevant range of
functioning. Even the best performing fit indices (CFI and TLI) were more variable and less accurate than parallel analysis—one of the most widely recommended methods in the dimensionality literature (Garrido et al., 2016; Horn, 1965). Subsequent studies by Clark and Bowles (2018) and Beierl, Bühner, and Heene (2018) have corroborated these findings, showing that none of the conventional cut-off values previously cited appeared to perform well enough to be recommended. Similar to Garrido et al. (2016), they found that the RMSEA index combined with the typical cut-off value of 0.05 was insensitive to latent misspecification, frequently accepting models with fewer major factors than those in the population. Similarly, while the CFI and TLI indices performed more reliably using a 0.95 cut-off value, their accuracy was contingent on the size of the factor loadings, factor correlations, and number of items (Beierl et al., 2018; Clark & Bowles, 2018). Again, the SRMR index performed very poorly, generally accepting underfactored models when using a cut-off value of 0.08.

In general, research studies evaluating the use of fit indices for dimensionality assessment have encountered similar problems to those found in the general literature that has assessed the usefulness of fit indices for latent variable modeling. Specifically, it has been shown that the values of the fit indices are affected by incidental parameters not related to the size of the misfit, such as the size of the factor loadings, the number of response categories, or the sample size (Chen et al., 2008; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Hu & Bentler, 1999; McNeish, An, & Hancock, 2018; Savalei, 2012; Xia & Yang, 2018). Because of this, there are no “golden rules” for fit indices, as no single cut-off value can be expected to perform well across the diverse type of data structures encountered in practice (Marsh, Hau, & Wen, 2004). In a recent study, Clark and Bowles (2018) concluded that, “the common fit statistics cut-offs employed in the literature are at best imperfect tools for guiding decisions regarding dimensionality when interpreting an item factor analysis” (Clark & Bowles, 2018, p. 555). Clark and Bowles (2018) suggest that fit indices should be used in conjunction with other methods for assessing dimensionality.

Exploratory Graph Analysis: An Overview

A recently developed dimensionality assessment technique, exploratory graph analysis (Golino & Epskamp, 2017), is part of a relatively new field of quantitative methods termed network psychometrics (Epskamp, Maris, Waldorp, & Borsboom, in press; Epskamp, Rhemtulla, & Borsboom, 2017). EGA estimates the dimensionality structure of multivariate data by combining network analysis with a community detection algorithm (i.e., a method to detect dimensions in networks), and has shown to be a promising dimensionality assessment technique in simulation studies (Golino & Epskamp, 2017; Golino et al., in press). Currently, however, there is no fit index that can directly determine the adequacy of the dimensional structures estimated via EGA (without having to use additional models such as CFA; Golino & Demetriou, 2017). This section briefly introduces EGA, which can be used as one of the steps for assessing dimensionality, complementing the suggestion of Clark and Bowles (2018).

EGA is a dimensionality technique from the network psychometrics perspective. A network is depicted by nodes (e.g., circles) representing variables and edges (e.g., lines) representing the associations between variables. The general notion behind the EGA algorithm is to estimate a network and apply the Walktrap community detection algorithm (Pons & Latapy, 2006), which is a common approach for
identifying communities (densely connected sets of nodes) in weighted networks. These communities are shown to be equivalent to latent factors of factor models (Golino & Epskamp, 2017; Golino et al., in press). Below we provide a more thorough explanation of this algorithm. Readers interested in a more in-depth description of EGA can find additional details in the simulation study and tutorial published by Golino et al. (in press).

The first network estimation method for EGA algorithm is the Gaussian graphical model (GGM: Lauritzen, 1996), which estimates multivariate joint probability distribution of random variables—that is, the probability for the values of each pair of variables to be within the range of values for those variables. This probability distribution is represented by the inverse of the covariance matrix (often refered to as the precision matrix), whose off-diagonal elements represent the conditional dependence between two variables given all other variables. The current approach used in network psychometrics to estimate a GGM works as follows. Let’s assume a set of random variables \( y \) that are normally distributed with mean zero and variance-covariance matrix \( \Sigma \), and let \( K \) (kappa) be the inverse of \( \Sigma \),

\[
K = \Sigma^{-1}
\]  

then, each element \( k_{ij} \) (row \( i \), column \( j \) of \( K \)) can be standardized to yield the partial correlation between two variables \( y_i \) and \( y_j \), given all other variables in \( y_{-c(i,j)} \) (Epskamp & Fried, 2018):

\[
\text{Cor}(Y_i, Y_j | y_{-c(i,j)}) = -\frac{k_{ij}}{\sqrt{k_{ii}k_{jj}}}
\]  

Epskamp and Fried (2018) points that modeling \( K \) in a way that every nonzero element is treated as a freely estimated parameter generates a sparse model for \( \Sigma \). The sparse model of the variance-covariance matrix is the GGM (Epskamp & Fried, 2018).

This partial correlation matrix represents a fully connected network where all nodes are connected to all other nodes. Typically, a network estimation method is applied to remove non-relevant relations resulting in a sparser network (i.e., each node is no longer connected to all other nodes). The level of sparsity of the GGM can be set using different methods, but the most common method is to apply a variant of the least absolute shrinkage and selection operator (LASSO; Tibshirani, 1996) termed graphical LASSO (GLASSO; Friedman, Hastie, & Tibshirani, 2008). The GLASSO is a regularization technique that estimates both the model structure and the parameters of a sparse GGM (Epskamp & Fried, 2018). It has a tuning parameter (\( \lambda \)) that is chosen by minimizing the extended Bayesian information criterion (EBIC; Chen & Chen, 2008), which has been shown to accurately retrieve the simulated network structure in Monte Carlo studies (Epskamp & Fried, 2018; Foygel & Drton, 2010).

A newer version of EGA (Golino et al., in press) uses a different network estimation method called triangulated maximally filtered graph method (Christensen, Cotter, & Silvia, 2019; TMFG; Massara et al., 2016). The TMFG method computes a zero-order
correlation matrix from the data and applies a constraint on the number of edges to retain in the network \((3n - 6, \text{where } n \text{ is the number of nodes; Massara et al., 2016})\). The algorithm begins by identifying the four variables with largest sum of correlations to all other variables and connects them to each other. Then, variables are iteratively added to the network based on the largest sum of three correlations to nodes already in the network. This process is repeated until all nodes are connected in the network. The resulting network is composed of 3- and 4-node cliques (i.e., sets of connected nodes; a triangle and tetrahedron, respectively). Hereafter, EGA will refer to EGA with GLASSO network estimation and EGAtmfg will refer to EGA with TMFG network.

Once the network is estimated using one of these two methods, the Walktrap community detection algorithm is applied. The Walktrap algorithm uses a process known as random walks or stochastic transitions from one node to another over an edge. To define the random walk, let \(A\) be a square matrix of edge weights (e.g., partial correlations) in the network, where \(A_{ij}\) is the strength of the (partial) correlation between node \(i\) and \(j\) and a node’s strength is the sum of node \(i\)’s connections to its neighbors \(NS = \sum_j A_{ij}\). The steps move from one node to another randomly and uniformly using a transition probability, \(P_{ij} = \frac{A_{ij}}{NS(i)}\), which forms the transition matrix, \(P\). To determine the communities that the nodes belong to, the transition matrix is used to compute a distance metric, \(r\), which measures the structural similarity between nodes (Equation 7). This structural similarity is defined as (Pons & Latapy, 2006):

\[
r_{ij} = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik} - P_{jk})^2}{NS(k)}}. \tag{7}
\]

This distance can be generalized to the distance between nodes and communities by beginning the random walk at a random node in a community, \(C\). This can be defined as:

\[
P_{Cj} = \frac{1}{|C|} \sum_{i \in C} P_{ij}. \tag{8}
\]

Finally, this can be further generalized to the distance between two communities:

\[
r_{C_1C_2} = \sqrt{\sum_{k=1}^{n} \frac{(P_{C_1k} - P_{C_2k})^2}{NS(k)}}, \tag{9}
\]

where this definition is consistent with the distance between nodes in the network (Equation 7).
The algorithm begins by having each node as a cluster (i.e., \( n \) clusters). The distances, \( r \), are computed between all adjacent nodes, and the algorithm then begins to iteratively choose two clusters. These two clusters chosen are then merged into a new cluster, updating the distances between the node(s) and cluster(s) with each merge (in each \( k = n - 1 \) steps).

Clusters are only merged if they are adjacent to one another (i.e., an edge between them). The merging method is based on Ward’s agglomerative clustering approach (Ward, 1963) that depends on the estimation of the squared distances between each node and its community \( (\sigma_k) \), for each \( k \) steps of the algorithm. Since computing \( \sigma_k \) is computationally expensive, Pons and Latapy (2006) adopted an efficient approximation that only depends on the nodes and the communities rather than the \( k \) steps. The approximation seeks to minimize the variation of \( \sigma \) that would be induced if two clusters \( (C_1, C_2) \) are merged into a new cluster \( (C_3) \):

\[
\Delta \sigma(C_1, C_2) = \frac{1}{n} \left( \sum_{i \in C_3} r_{iC_3}^2 - \sum_{i \in C_1} r_{iC_1}^2 - \sum_{i \in C_2} r_{iC_2}^2 \right).
\]

Since Ward’s approximation adopted by Pons and Latapy (2006) only merges adjacent clusters, the total number of times \( \Delta \sigma \) is updated is not very large, and the resulting values can be stored in a balanced tree. A sequence of \( P_k \) partitions into clusters \( (1 \leq k \leq n \), being \( n \) the total number of nodes) is obtained. The best number of clusters is defined as the partition that maximizes modularity.

Modularity is a measure that was proposed by Newman (2003) to identify meaningful clusters in networks and is calculated as follows. Let \( j \) and \( k \) be two clusters in a network with \( m \) and \( n \) nodes. If the number of edges between clusters is \( p \), then one-half of fraction of the edges linking \( j \) and \( k \) is \( e_{jk} = \frac{1}{2} p \), so that the total fraction of edges between the two clusters is \( e_{jk} + e_{kj} \) (Newman, 2003). Conversely, \( e_{jj} \) represents the fraction of edges that fall within cluster \( j \), whose sum equals one: \( \sum_j e_{jj} = 1 \). Newman (2003) points out that a division of networks into clusters is meaningful if the value of the sums of \( e_{jj} \) and \( e_{kk} \) is maximized. However, in cases where only one cluster is presented, the maximal value will be one, which is also the value of \( \sum_j e_{jj} \). Therefore, for networks composed by only one cluster this index is not informative. A solution Newman (2003) proposed was to calculate an index that takes \( \sum_j e_{jj} \) and subtract from it the value that it would take if edges were placed at random. For a given cluster \( j \), the modularity is calculated as:

\[
Q = \sum_j \left( e_{jj} - a_j^2 \right),
\]

where \( a_j \) is given by \( \sum_i e_{jk} \). Therefore, the modularity index penalizes network structures with only one cluster, since in this condition the value of \( Q \) would be zero (Newman, 2003).

Golino and Epskamp (2017) studied the accuracy in estimating the number of dimensions of EGA along with six classical techniques in a simulation manipulating
the number of factors, number of items, sample size, and correlation between factors. Golino and Epskamp (2017) showed that the most accurate techniques were PA (89% of accurate estimates) and EGA (93%). More recently, Golino et al. (in press) compared EGA and EGAtmfg along with five factor-analytic techniques, including two automated versions of Cattell’s scree test (Cattell, 1966): The scree test optimal coordinate (OC) and acceleration factor (AF) methods (Raiche, Walls, Magis, Riopel, & Blais, 2013). The results presented by Golino et al. (in press) showed that EGA (87.91%), parallel analysis (83.01%) and EGAtmfg (74.61%) presented high to moderately high accurate estimations (i.e. correctly estimating the number of simulated factors). Golino et al. (in press) also showed that EGA works well for both dichotomous (overall accuracy of 85%) and continuous data (overall accuracy of 91%), although the accuracy in estimating the number of simulated factors is higher in the latter type of data.

On the one hand, Golino et al. (in press) found that when EGA and EGAtmfg provided the same estimate for 78% of the datasets, they had a nearly perfect accuracy (91.85%). On the other hand, when the two methods disagreed, the accuracy of both EGA (73.3%) and EGAtmfg (12.9%) diminished. In these cases where EGA and EGAtmfg disagree, it’s important to consider potential alternative solutions (with more or less dimensions) to those suggested by the methods. In this case, a fit index specifically designed for dimensionality assessment would play a crucial role, motivating the current article.

Another issue of central importance is the number of steps used in the random walk process of the Walktrap algorithm. To date, EGA has been evaluated using the default value of four steps. Given previous simulation evidence, four steps appears to be adequate for the number of variables typically used in psychological constructs, which has been verified in simulations outside of the area (Pons & Latapy, 2006). Indeed, in Pons and Latapy’s (2006) seminal study, three and four steps appeared to be optimal when there are fewer than 100 variables. Selecting the optimal number of steps is important since it may increase the precision of the EGA estimation, leading to more reliable and interpretable factors. Therefore, developing a fit metric that could be used to compare the dimensionality structure estimated via EGA with different number of steps used in the walktrap algorithm (e.g. from 3 to 10), could improve the quality of the EGA procedure.

### Entropy and the Structure of Multivariate Data

The current paper presents a family of three new fit indices, *Entropy Fit Indices*, which are based on an information-theoretic concept of entropy that can be used to check the fit of the dimensionality structures estimated via EGA and other factor-analytic methods. The next paragraphs will introduce the main concepts used to develop the EFIs, starting with the concept of Shannon’s entropy (Shannon, 1948) and later introducing Von Neumann’s entropy (Von Neumann, 1927). The goal is to help the reader understand the conceptual and technical origins of the entropy fit measures proposed in the current article.
Shannon’s Entropy: Estimation and Properties

The study of entropy originated in physics, more specifically in thermodynamics, with the work of Boltzmann and others (Watanabe, 1960). Although a very specific type of entropy was being used in communication theory as early as 1928 (Watanabe, 1960), a formal mathematical theory of communication based on entropy was later developed by Shannon (1948). The use of entropy has increased significantly over the years in many different areas such as communication, neuroscience (Paninski, 2003), biology (Tsuruyama, 2018), ecology (Harte & Newman, 2014), physics (Zurek, 2018), chemistry (Ferenci & Kovács, 2014), and other fields (Volkenstein, 2009).

The entropy of a discrete random variable \( X \) characterizes the uncertainty associated with \( X \) or the degree of disorganization of \( X \) (Wiener, 1961). In other words, it depicts the average amount of unpredictability that is removed when the outcome of \( X \) is known (Yeung, 2008), which can also be considered a measure of information of \( X \) (entropy and information are negatively related). Entropy depends on the distribution of \( p(X) \), and is maximized when all events of \( X \) are equiprobable (i.e., \( p(x) = \frac{1}{N} \), where \( N \) is the number of elements of \( X \); Shannon, 1948).

The maximum likelihood estimator (MLE) or the plug-in method (Antos & Kontoyiannis, 2001) is one of the most common methods to estimate the entropy of a random variable. Let \( X \) be a discrete random variable that takes on values in a set \( \mathcal{X} = (x_1, x_2, ..., x_n) \), with a probability mass function \( p(x) = P(X = x) \). Antos and Kontoyiannis (2001) showed that considering additive functionals with the general form:

\[
F = \sum_{x \in \mathcal{X}} f(x, p(x)),
\]

where \( f \) and \( g \) are non-negative arbitrary real-valued functions, the plug-in estimate for \( F \) is \( \hat{F} = F(p_n) \), where \( p_n \) is the empirical distribution:

\[
p_n(x) = \frac{1}{n} \sum_{j=1}^{n} I(X_j = x).
\]

The MLE of entropy is, thus:

\[
\hat{H}(p_n) = - \sum_{i=1}^{n} p_{n,x} \log p_{n,x}.
\]

Or, more directly:
\[ H(X) = - \sum_{i=1}^{n} p(x) \log p(x). \]  

(15)

Entropy can also be extended to multiple variables. Let \( X \) and \( Y \) be discrete random variables, with a joint probability of \( p(x, y) \). The entropy of the joint distribution is (Shannon, 1948):

\[ H(X, Y) = - \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y). \]  

(16)

In the case where \( Y \) and \( X \) are independent variables, their joint entropy will always be less than or equal to the sum of their individual entropies:

\[ H(X, Y) \leq H(X) + H(Y). \]  

(17)

If \( X \) and \( Y \) are not independent variables, presenting a conditional probability distribution for \( Y \) given \( X \) as \( p(y|x) \), then the conditional entropy \( H(Y|X) \) is “the average of the entropy of \( Y \), for each value of \( X \), weighted according to the probability of getting that particular \( X \)” (Shannon, 1948, p. 12). Mezard and Montanari (2009) summarizes the idea of conditional entropy as the amount of information we obtain from the value of \( y \) if one knows \( x \). The MLE estimation of conditional entropy is:

\[ H(Y|X) = - \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log p(y|x). \]  

(18)

The joint entropy of \( X \) and \( Y \) can be re-written as the entropy of \( X \) plus the entropy of \( Y \) given \( X \):

\[ H(X,Y) = H(X) + H(Y|X). \]  

(19)

In other words, the uncertainty of the joint event \((X,Y)\) is the uncertainty of \( X \) plus the uncertainty of \( Y \) when \( X \) is known (Shannon, 1948). Hence, the uncertainty of \( Y \) never increases by knowledge of \( X \):

\[ H(X) + H(Y) \geq H(X,Y) = H(X) + H(Y|X) \therefore H(Y) \geq H(Y|X). \]  

(20)

Also, if \( X \) and \( Y \) are independent, then \( H(Y) = H(Y|X) \).
In his seminal paper, Watanabe (1939) created a modified index of entropy termed total correlation to measure not only the uncertainty of random variables, but also the strength of their correlation beyond their average interaction (Watanabe, 1960). The total correlation of a set of variables $\mathcal{X} = (x_1, x_2, ..., x_n)$ is a non-negative index that increases as entropy of individual variables increases or as entropy of the set of variables (the joint entropy of $\mathcal{X}$) decreases (Watanabe, 1960), and is calculated as follows:

$$C_{\text{tot}, \mathcal{X}} = \left( \sum_{i=1}^{n} H(x_i) \right) - H(x_1, x_2, ..., x_n) \geq 0. \quad (21)$$

The interpretation of total correlation provided by Watanabe (1960) can be illustrated with a very simple example. Suppose the set of variables $\mathcal{X}$ is divided in two subsets: $\nu$, containing half the elements of $\mathcal{X}$; and $\omega$ containing the other half. Observing the variables contained in $\nu$ and $\omega$ separately, the information in each subset is $H(\nu)$ and $H(\omega)$. The individual information in each subset of variables individually is higher than the information carried by $\mathcal{X}$ if and only if $\nu$ and $\omega$ are independent. In this scenario of independence, total correlation is the loss of information or redundancy. If $H(\nu)$ and $H(\omega)$ are dependent, then the difference between their joint entropy and the sum of their individual entropies is a measure of the correlation between them, reflecting ignorance before observation. Watanabe (1960) point out that without any observation, the ignorance about a subset $\nu$ is $H(X)$, but after observing $\nu$ the ignorance becomes $H(X) - H(\nu)$. The decrease of ignorance is thus the information about $\omega$ provided by the observation of $\nu$.

Total correlation was used as the basis for a new technique termed interdependence analysis (IDA; Watanabe, 1969), that replaced a metric of distance to this entropic measure of correlation in cluster identification. The general idea of interdependence analysis is to partition a set into subsets until interdependence is minimized (Watanabe, 2000). Watanabe (2000) showed that the construction of clusters using total correlation could lead to unwanted problems in the partitioning of a multidimensional space of variables. In particular, by not considering the size of the subsets of a multidimensional space, small subsets could be removed due to their small entropy value, irrespective of the relationship to the other subsets. Watanabe (2000, 2001), introduced a slight modification to the IDA process by modifying the total correlation index (Equation 21), replacing it with the $K$-function. Considering a set of variables $\mathcal{X} = (x_1, x_2, ..., x_n)$, and $X_{\nu}$ and $X_{\omega}$, with cardinalities $n_1$ and $n_2$, respectively, the $K$-function representing the interdependence between $X_{\nu}$ and $X_{\omega}$ is given by:

$$k(X_{\nu}, X_{\omega}) = n_1 H(X_{\nu}) + n_2 H(X_{\omega}) - (n_1 + n_2)H(X_{\nu}, X_{\omega}). \quad (22)$$

Contrary to the entropy index that is infra-additive (see Equation 17), the product of entropy by the cardinalities is supra-additive, making the $K$-function non-positive:
Using the notation above, the $K$-function is the difference between the average entropy of $X_v$ and $X_\omega$ from the entropy of the super-set $\mathcal{X}$. The $K$-function can be used as a measure of entropy reduction by partitioning, in which the minimization is the minimization of the uncertainty or unstableness (Watanabe, 2000, 2001) in the partition of a multidimensional space. It can also be interpreted as the maximization of information gain by the partitioning of a multidimensional space (Watanabe, 2000). Despite its interesting characteristics, the $K$-function could not be used as a fit index for dimensionality assessment (or for clustering) because it decreases with the increase in the number of subgroups of variables (or clusters). So, the application of $K$-function to decide on the number of clusters, factors or partitionings of a multidimensional space is limited.

The Entropy Fit Indices: Properties and Estimation

Given the limitation of the $K$-function proposed by Watanabe (2000), it is necessary to develop a new metric that does not decreases with the increase in the number of partitions of a multidimensional space (or clusters, factors and so on). The Entropy Fit Index (EFI) was developed as an alternative to the $K$-function, with the goal to identify the best structure of a given set of random variables, when comparing two or more alternative structures. In other words, EFI can be used to identify the best partitioning of a multidimensional space that better reflects the underlying latent factors, without necessarily decreasing with the increase in the number of factors.

Let $\eta$ be a set with two latent factors ($\eta = (\eta_1, \eta_2)$) with variance-covariance matrix $\Sigma$:

$$\Sigma = \Lambda \Phi \Lambda' + \Theta,$$

(24)

where $\lambda$ is the measurement model (i.e., a $k \times 2$ factor loading matrix for $k$ variables and two factors), $\phi$ is the structure matrix of the latent variables (i.e., a $2 \times 2$ matrix of correlations among factors), and $\theta$ is the covariance matrix of the residuals. Now, let $\eta_1$ and $\eta_2$ be two latent factors with observable variables $s_1, s_2, ..., s_n$ and $t_1, t_2, ..., t_m$, respectively. The sum scores of the variables in each factor is:

$$S_{\eta_i} = \sum_{i=1}^{n} s_i$$

(25)

and
\[ S_{\eta_2} = \sum_{i=1}^{m} t_i. \]  

(26)

The entropy fit index of the structure formed by \( \eta_1 \) and \( \eta_2 \) can now be computed as:

\[
\text{EFI} = \left[ \frac{H(S_{\eta_1}) + H(S_{\eta_2})}{2} - H(S_{\eta_1}, S_{\eta_2}) \right] + \left[ \left( H_{\text{max}} - \frac{H(S_{\eta_1}) + H(S_{\eta_2})}{2} \right) \times \sqrt{2} \right].
\]

(27)

where \( H(S_{\eta_1}) \) and \( H(S_{\eta_2}) \) are the entropy of the sum scores of the variables from factors \( \eta_1 \) and \( \eta_2 \), respectively, \( H(S_{\eta_1}, S_{\eta_2}) \) is the joint entropy of the sum scores, and 2 is the number of factors. The entropy of the sum scores for all variables from \( \eta_1 \) and \( \eta_2 \) is \( H_{\text{max}} \). So, \( H_{\text{max}} \) can be considered as the entropy estimated for a single factor containing all the \( n + m \) variables in \( \eta_1 \) and \( \eta_2 \). Finally, the difference between the maximum entropy (i.e., entropy estimated for a single factor containing all the \( n + m \) variables) and the average entropy of the factors is multiplied by the square root of the number of factors (from now on this difference will be termed difference entropy).

In sum, Equation 27 calculates the difference between the average entropy for the individual factors and the joint entropy of the factors, while including a penalization for the number of factors. The penalization modulates the EFI to not decrease with an increase in the number of factors, except if each additional factor decreases the uncertainty or disorder of the system of variables. Equation 27 can be easily extended to \( N_F \) factors:

\[
\text{EFI} = \left[ \frac{\sum_{i=1}^{N_F} H(S_{\eta_i})}{N_F} - H(S_{\eta_1}, S_{\eta_2}, \ldots, S_{\eta_N}) \right] + \left[ \left( H_{\text{max}} - \frac{\sum_{i=1}^{N_F} H(S_{\eta_i})}{N_F} \right) \times \sqrt{N_F} \right].
\]

(28)

The EFI can be used as a measure of entropy reduction by partitioning of a multidimensional space into several groups of variables, in which the minimization is the minimization of the uncertainty or disorder due to the correct identification of the underlying factors. In contrast with the K-function, EFI does not decrease with the increase in the number of factors. The EFI will be lower if the correct structure is provided (i.e., the correct number of factors and the correct composition of items per factor), irrespective of the number of factors. This principle is due to the entropy and the fineness of classification (e.g. classifying items into subgroups) theorem (Watanabe, 2000) that states the following:

**Theorem 1 (Entropy and the Fineness of Classification).** Let \( X_I \) and \( X_J \) be subsets of \( X \), and \( A \) be a fixed set of predicates. If a classification of \( A \) with respect to the latter is a refinement of the classification with respect to the former then the formal entropy of \( X_I \) is never larger than that of \( X_J \).

Theorem 1 implies that a structure that splits a set of random variables into groups, which reflect the underlying latent factors, will have a lower entropy than a structure
that does not reflect the latent factors. An example helps illustrate this principle. Let \( \eta_1 \) and \( \eta_2 \) be two latent factors with four variables each. For every unique combination of the items into two sets of four, the EFI is recorded, generating the result showed in Figure 1. The first and last elements of the x-axis represent the correct structure \((1, 1, 1, 1, 2, 2, 2, 2)\) or \((2, 2, 2, 2, 1, 1, 1, 1)\), while the other elements of the x-axis presents all the other possible combinations of the variables into two factors (e.g., \(1, 1, 1, 1, 2, 2, 2, 2\)). At the bottom of Figure 1, a matrix representing the distribution of the variables into two factors (black representing variables from factor \( \eta_1 \), white representing variables from factor \( \eta_2 \)) is shown. The columns of the matrix shows the distribution of the items into two groups, while the rows are divided into two parts, the first four rows represents \( \eta_1 \), the last four rows represents \( \eta_2 \).

Figure 1 shows a perfect symmetric relationship between the distribution of variables per factor and the EFI index (y-axis). The EFI is the same for the structure \((1, 1, 1, 1, 2, 2, 2, 2)\) and \((2, 2, 2, 2, 1, 1, 1, 1)\). Also, \((1, 1, 1, 1, 2, 2, 2, 2)\), second element in the x-axis, and \((2, 2, 2, 2, 1, 1, 1, 1)\), 69th element of the x-axis, has the same EFI. This symmetric behavior of EFI can be explained by the nature of the entropy measures based on total correlation or on the \( K \)-function: it reflects the relationship among groups of variables rather than the relationship between variables. Thus, every time a group of variables is classified in an incorrect way (i.e., not reflecting the true underlying latent factors) there is loss of information and the EFI increases. For example, if EFI were computed in the same data used to generate Figure 1, but with a three-factor structure instead of two (i.e., \(1, 1, 1, 2, 2, 3, 3\)), the EFI would increase from \(-0.402\) to \(-0.328\).

It is important to note that Figure 1 focuses on a multidimensional structure with two
factors and all combinations of variables in two groups of four variables. The pattern is different if we vary the number of factors, in which the EFI value would be lower only if the correct structure is provided, increasing linearly with the number of factors. This is an important difference between EFI and the $K$-function, since the latter always decreases with the increase in the number of groups of variables (or factors).

To illustrate this principle, consider a multidimensional structure with 100 factors, four variables per factor and a process with 100 steps. In the first steps, the EFI, total correlation, $K$-function, and the difference between $H_{max}$ and the average entropy of the factors ($\text{difference entropy}$) is recorded using the correct structure (i.e., 100 factors, four variables per factor). In each subsequent step one factor is split into two sub-factors, with two variables each and all the indexes used in the first step are recorded. The result of this process shows the variability of the EFI, total correlation, $K$-function, and difference entropy as the factors are sequentially sub-divided into two factors.

![Graphs showing the change in entropy-related measures as factors and variables are divided.](image)

**Figure 2.** Entropy fit index per combination of variables into two factors

Figure 2 shows exactly the pattern described above. The $K$-Function of the factors (or average entropy of the factors) decreasing with the number of latent variables, while the difference entropy increases following an exponential function that can be approximated by the following formula:

$$H_{Dif} \simeq 1.1568 \times \frac{e^{(0.0335 \times N_F)}}{1 + e^{(0.0335 \times N_F)}}.$$  \hspace{1em} (29)
In sum, what Figure 2 shows is that how EFI differs from the *K-function*, being more suitable as a measure of fit due to the difference entropy, which penalizes the number of partitions of the multidimensional space or number of factors.

In terms of estimation, several estimators are available for entropy (especially empirical entropy or Shannon entropy; Paninski, 2003). In the current article, the maximum likelihood estimator of entropy (Equation 15) is used to calculate the EFI. In the MLE of entropy, the number of bins to compute the empirical distribution of the continuous variables needs to be provided. The optimal number of bins can be estimated using the formula suggested by Cellucci, Albano, and Rapp (2005), where the number of bins is the greatest integer such that:

\[ N_{Bins} \leq \sqrt{\frac{N}{5}}, \]  

where \( N \) is the sample size. So, after calculating the number of bins, the sum score of each group of variables (indicating a specific latent factor) is partitioned into \( N_{Bins} \) elements, and Equation 28 is computed. It is important to note that the plug-in estimator of entropy, as used in the current paper, can only be computed with raw data.

**Entropy Fit Index with Von Neumann’s Entropy and the Total Entropy Index: Estimations using a correlation matrix**

Because in the areas of psychology, education, health and related fields it is common for a researcher to have access only to the correlation matrix of multivariate datasets, it is valuable to have an alternative estimation for the EFI that could be computed using correlations. A plausible and approximate solution is to use matrix algebra to calculate Von Neumann entropy (Von Neumann, 1927), an index that was developed to quantify the amount of disorder in a system, as well as the entanglement between two subsystems in quantum physics (Preskill, 2016), which occurs when two (or more) particles become inextricably linked. The entanglement of two photons can be verified using specific visualizations of the spatial correlations and momentum correlations (Moreau et al., 2019). In this case, the visualization provides a clue to how linked (entangled or mixed) two systems (or particles) are. The entanglement of two or more systems can also be expressed in terms of a density matrix, a type of matrix used in quantum mechanics that is akin to the probability distribution of position and momentum (i.e. phase-state probability) in classical statistical mechanics (Hall, 2013). Any density matrix has three main characteristics: (1) it is real symmetric, (2) positive semi-definite, and (3) has trace equal to one.

Preskill (2016) shows that for any density matrix \( \rho \), Von Neumann entropy is:

\[ S(\rho) = -tr(\rho \log \rho). \]  

Any correlation matrix can be transformed into a density-like matrix by using the number of variables as a scaling factor that will make the trace of the matrix equal to...
one (Anderson & others, 1963). By dividing the elements of the matrix by the number of variables, the correlation matrix will hold the same properties as any density matrix.

Wihler, Bessire, and Stefanov (2014) shows that eigenvalues can be used to estimate Von Neumann entropy. Given a density matrix \( \rho \) with eigenvalues \( \lambda_1, \lambda_2, ..., \lambda_m \geq 0 \), Von Neumann entropy is:

\[
S(\rho) = -\text{tr}(\mathcal{L}(D)),
\]

where \( \mathcal{L} \) is a continuous function on the real interval \([0, \infty)\) that takes values \( x \log(x) \) if \( x \geq 0 \) and 0 if \( x = 0 \), and \( D = (\lambda_1, \lambda_2, ..., \lambda_m) \). In other words, the Von Neumann entropy of a density matrix is the Shannon entropy of the vector of eigenvalues (Preskill, 2016).

As presented earlier, to calculate the entropy fit index of a given structure using Shannon’s entropy the joint entropy of the factors must be computed. The joint Von Neumann entropy of two (or more) density matrices can be calculated as the Kronecker product of the matrices. Given two density matrices, \( \rho_A \) and \( \rho_B \), the product state \( \rho_{AB} \) is given by \( \rho_A \otimes \rho_B \) (Preskill, 2016). Therefore, after calculating \( \rho_{AB} \), Equation 32 can be used to estimate the Von Neumann entropy of the joint state (or the joint Von Neumann entropy). If the number of joint states is large (e.g., more than five), the resulting Kronecker product can be very large. To avoid this issue, the product of the individual eigenvalues can be used instead of the Kronecker products, since the eigenvalues of \( A \otimes B \) is \( \lambda_1 \mu_1, ..., \lambda_1 \mu_m, ..., \lambda_2 \mu_1, ..., \lambda_2 \mu_m, ..., \lambda_n \mu_m \) (Laub, 2005).

Since the EFI is estimated using the raw data (i.e., Shannon’s entropy), while Von Neumann entropy is estimated using a density-like, scaled version of a correlation matrix, two modifications from Equation 28 are necessary to estimate obtain the EFI Von Neumann (EFI.vn; see Equation 33). First, the sum scores of variables belonging to the factors are replaced by sub-matrices of \( \rho \). So, two factors can be represented as \( \rho_1 \) and \( \rho_2 \), each containing a different combination of rows and columns of \( \rho \). If the first factor is composed by variables \( i, j \) and \( k \) of \( \rho \), the sub-matrix \( \rho_1 \) will be:

\[
\rho_1 = \begin{bmatrix}
i i & i j & i k \\
i j & j j & j k \\
i k & j k & k k
\end{bmatrix}
\]

The second factor being composed by variables \( m, n \) and \( o \) of \( \rho \), the sub-matrix \( \rho_2 \) will be:

\[
\rho_2 = \begin{bmatrix}
m m & m n & m o \\
n m & n n & n o \\
o m & n o & o o
\end{bmatrix}
\]

The Von Neumann entropy of \( \rho_1 \) and \( \rho_2 \) are \( S(\rho_1) \) and \( S(\rho_2) \), respectively.
The second modification relates to the penalization, in which the maximum entropy is replaced by the total entropy of the matrix $S(\rho)$. Also, the penalization now is subtracted, instead of added as in Equation 28. EFI.vn can be computed for two factors as follows:

$$EFI_{VN} = \left[ \frac{S(\rho_1) + S(\rho_2)}{2} - S(\rho_1 \otimes \rho_2) \right] - \left[ \left( S(\rho) - \frac{S(\rho_2) + S(\rho_1)}{2} \right) \times \sqrt{2} \right].$$

(33)

where $S(\rho_1)$ and $S(\rho_2)$ are the Von Neumann entropies for factors one and two, $S(\rho)$ is the total entropy of the matrix $\rho$ and $S(\rho_1 \otimes \rho_1)$ is the joint entropy of the factors, which can be calculated using the Kronecker product of the matrices or (in case of multiple factors) as the product of eigenvalues, as pointed out earlier. Equation 33 can be easily expanded to $N_F$ factors:

$$EFI_{VN} = \left[ \frac{\sum_{i=1}^{N_F} S(\rho_i)}{N_F} - S(\rho_1 \otimes \rho_2 \otimes \ldots \otimes \rho_i) \right] - \left[ \left( S(\rho) - \frac{\sum_{i=1}^{N_F} S(\rho_i)}{N_F} \right) \times \sqrt{N_F} \right].$$

(34)

The use of the joint entropy of the factors can be a possible limitation for EFI.vn. While in the EFI the joint entropy is estimated directly using the raw data, in the EFI.vn is calculated using the Kronecker product of the individual factors (or the product of the individual eigenvalues per factor, as shown above). This strategy seems useful to detect superposition of states (in the present case the superposition is when two factors are merged into one), since this is a typical property of quantum information theory (Preskill, 2016; Zhao, Yang, Hung, & Li, 2015). However, if one pure state (one factor) is split into two, generating a plausible but incorrect structure, then the performance of the EFI.vn may not be very accurate.

An alternative approach to EFI.vn is to use the total entropy in the correlation matrix, which can be calculated using Equation 32, instead of the joint entropy of the factors in the left side of Equation 33. This alternative EFI.vn measure, hereafter named total entropy fit index with Von Neumman entropy (TEFI.vn), maintains the same properties as EFI and EFI.vn by modifying the penalization using only the sum of the factor entropies instead of the mean. Then, TEFI.vn can be calculated as:

$$TEFI_{VN} = \left[ \frac{\sum_{i=1}^{N_F} S(\rho_i)}{N_F} - S(\rho) \right] + \left[ \left( S(\rho) - \sum_{i=1}^{N_F} S(\rho_i) \right) \times \sqrt{N_F} \right].$$

(35)

Methods

Design

To investigate the adequacy of the entropy fit indices (i.e., EFI, EFI.vn and TEFI.vn) and the traditional fit indices used in factor models (CFI, RMSEA, TLI and SRMR) two Monte-Carlo simulations are presented. The first simulation identifies the best
fit indices under the assumption that the correct number of factors is known. In other words, given two latent factors with four variables each, which fit index is the most accurate in identifying the correct placement of variables within the two factors? To answer this question, a permutation strategy is implemented, similar to the first example showed in the last section. The fit indices are computed for every unique combination of the items into two sets of four (in a total of 70 unique combinations, from \((1,1,1,2,2,2,2,2,2,2,2,2)\), \((1,1,2,1,2,2,2,1,1,1,1,1)\), \((2,2,2,1,2,1,1,1,1,1,1,1)\). Four between-subject data factors were systematically manipulated: the skew of the items \((-2; 2; -2; 2; -2; 2; -2; 2\) in increments of 0.5), sample size (1000 and 5000), factor loadings (.40, .55 and .70) and factor correlations (0, .5 and .7).

In the second Monte-Carlo simulation, the fit indices will be compared in terms of their capacity to identify the correct structure (number of factors and the correct placement of items per factor) in four different set of conditions (shuffled, underfactoring, overfactoring with mixed items and overfactoring with a more extreme condition; i.e., no mixed items). For each set of conditions, the fit index for the correct structure (i.e., the structure that corresponds exactly to the simulated population structure; see Figure 3.A and Figure 4.A) is compared to an incorrect structure. The incorrect structures were set as follows: One structure presents the same number of factors as the population structure, but with a random distribution or placement of items per factor (shuffled condition; Figure 3.B and Figure 4.B). A second structure reflects the number of factors in the population minus one (underfactoring condition; Figure 3.C and Figure 4.C), while a third organization presents the same underfactoring structure, but with a random assignment of items per factor (underfactoring condition with random placement of items; Figure 3.D and Figure 4.D).

The fifth and sixth structures reflect the number of factors in the population plus one (overfactoring condition with mixed items; Figure 3.E and Figure 4.E), with only one pure factor (containing items from one single factor) and a randomly assignment of items per factor overfactoring condition with random placement of items; Figure 3.F and Figure 4.F). The last condition reflects the number of factors in the population plus one with a more extreme condition where the factors contain items from one single dimension (overfactoring condition with a more extreme condition; i.e., no mixed items; Figure 3.G and Figure 4.G).

Finally, four between-subject data factors were systematically manipulated in the second simulation: factor loadings (.40, .55 and .70), factor correlations (0, 0.5 and 0.7), number of items per factor (4 and 8) and sample size (1000 and 5000). The number of factors was held constant (i.e., three factors).

Factor loadings of .40, .55 and .70 can be considered as poor, good, and excellent, respectively; thus, representing a wide range of factor saturations (Comrey & Lee, 2013). The factor correlations simulated include the orthogonal condition (.00) and large (.50) correlation levels, according to Cohen (1988), and very large factor correlations of .70. The reason to simulate data with very high factor correlations (.70) is because in some areas within psychology (e.g., intelligence), researchers sometimes have to distinguish between constructs that are highly correlated (e.g., Kane, Hambrick, & Conway, 2005). The factors generated were composed of four and eight indicators with salient loadings, following previous simulation studies in dimensionality assessment (Golino & Epskamp, 2017; Golino et al., in press; Velicer, 1976; Widaman, 1993). Sample sizes 1,000 and 5,000 can be considered large (Li, 2016) and very large, the latter allowing for the evaluation of the dimensionality methods.
Figure 3. Population structure and examples of models tested, 4 items per factor
Figure 4. Population structure and examples of models tested, 8 items per factor
in conditions that can approximate their population performance. Additionally, these sample sizes were selected by taking into account that tetrachoric correlations require large sample sizes to achieve acceptable sampling errors (Timmerman & Lorenzo-Seva, 2011).

Data Generation

For each part of the simulation, 500 sample data matrices of binary variables were generated according to a common factor model procedure that works as follows. First, the reproduced population correlation matrix (with communalities in the diagonal) was computed:

\[
R_R = \Lambda \Phi \Lambda',
\]

(36)

where \(R_R\) is the reproduced population correlation matrix, \(\Lambda\) is the measurement model (i.e., a \(k \times r\) factor loading matrix for \(k\) variables and \(r\) factors), and \(\Phi\) is the structure matrix of the latent variables (i.e., a \(r \times r\) matrix of correlations among factors). This procedure implies that the generated data does not contain correlated residuals (minor factors) at the population level.

The population correlation matrix \(R_P\) was then obtained by inserting unities in the diagonal of \(R_R\), thereby raising the matrix to full rank. The next step was performing a Cholesky decomposition of \(R_P\), such that:

\[
R_P = U'U.
\]

(37)

If either \(R_P\) was not semi-positive definite (i.e., at least one eigenvalue was \(\leq 0\)) or an item’s communality was greater than 0.90, the \(\Lambda\) matrix was replaced and a new \(R_P\) matrix was computed following the same procedure. Subsequently, the sample data matrix of continuous variables was computed as:

\[
X = ZU,
\]

(38)

where \(Z\) is a matrix of random standard normal deviates with rows equal to the sample size and columns equal to the number of variables.

The resulting continuous variables were dichotomized by applying a set of thresholds according to specific levels of skewness. In the first part of the simulation (the permutation of items per factors), the following levels of skewness were used: \(-2, 2, -2\) and 2, and from \(-2\) to 2 in increments of 0.5. In the second part of the simulation, because binary variables are not necessarily symmetrically distributed in practice, the level of skewness of each item was randomly assigned with equal probability from a set of possible values that ranged from \(-2\) to 2 in steps of 0.50. The thresholds used
to generate these levels of skewness for the binary variables were taken from (Garrido, Abad, & Ponsoda, 2011; Garrido et al., 2013).

It is very common in practice to find complex structures in which items present non-zero loadings in multiple factors. To generate cross-loadings with magnitudes consistent to those commonly found in real data (Bollmann, Heene, Küchenhoff, & Bühner, 2015), the procedure described in Meade (2008) was followed for the second part of the simulation: cross-loadings were randomly drawn from a normal distribution (with mean zero and variance of .05) for all the items except for the first two in each factor, which were set as markers (i.e., all of their cross-loadings were fixed to zero).

Of note, regarding the generation of the main loadings: The function generates the main loadings by drawing random values from a uniform distribution that has a range of ±.10 from the specified value. For example, if the main loadings are set at 0.70, the function generates loading values between 0.60 and 0.80.

Data Analysis

We used R (R Core Team, 2018) for all our analyses. The EFI, EFI.vn, and TEFI.vn were computed using the EGAnet package (Golino & Christensen, 2019), while CFI, RMSEA, TLI and SRMR were calculated using lavaan (Rosseel, 2012). Confirmatory factor models were estimated using the WLSMV estimator. The choice to use the WLSMV estimator was due to the discrete nature of the simulated data (dichotomous), since this estimator is theoretically justified for the factor analysis of discrete data and it gives the best results in simulation studies (Barendse et al., 2015; Beauducel & Herzberg, 2006). The figures were generated using the ggplot2 package (Wickham, 2016) and the ggpubr package (Kassambara, 2018). To estimate the EFI.vn TEFI.vn, a tetrachoric correlation was computed using the qgraph package (Epskamp, Cramer, Waldorp, Schmittmann, & Borsboom, 2012).

In order to evaluate the performance of the fit indices, the percentage of correct structure selection (PC) was used:

\[
PC = \frac{\sum C}{N}, \quad \text{for} \quad C = \begin{cases} 
1 \text{ if } S_{\text{selected}} = S_{\text{correct}} \\
0 \text{ if } S_{\text{selected}} \neq S_{\text{correct}} 
\end{cases}
\]

(39)

where \( N \) is the number of sample data matrices simulated, \( S_{\text{selected}} \) is the selected structure and \( S_{\text{correct}} \) is the correct structure. The PC criterion has boundaries of 0% and 100%, with 0% signaling complete inaccuracy and 100% indicating perfect accuracy. Since in the current simulation only one structure out of seven is correct, PC has a theoretical base rate near 14%.

The strategies used to select a structure depend on the fit index used. For EFI and EFI.vn, the selected structure is always the one presenting the lower value. For the other indices, two main strategies were used. The first one is based on the cut-off values only: CFI and TLI ≥ .95, RMSEA and SRMR ≤ .05. If a structure other than the correct structure presented a value that is considered adequate (passes the cut-off points), then the fit index received a score of zero (i.e., failed to identify the correct structure). The second strategy uses the fit indices as relative measures of fit. Considering all the conditions where the correct structure presented an adequate fit
value, the structure with the highest (CFI and TLI) or lowest (RMSEA and SRMR) fit values is selected. If the selected structure is not the correct one, the respective score is zero. Scores of one are only achieved when the selected structure is the correct structure. From now on, when the traditional fit indices are used as relative measures of fit they will be referenced using the abbreviation .rel: CFI.rel, TLI.rel, RMSEA.rel and SRMR.rel.

For reproducibility purposes, all R scripts used in the current article, as well as the Rmarkdown file with the manuscript and codes used in the analysis of the result, are available in an online repository at the Open Science Framework platform here.

Results

Combination of variables into two factors

The results of the first simulation shows that the fit indices can be separated into two groups: one with high PC (TEFI.vn, EFI.vn, SRMR.rel, EFI, TLI.rel, RMSEA.rel and CFI.rel) and the other with low PC (TLI, CFI, RMSEA, SRMR). The most accurate fit indices, from best to worst, were: TEFI.vn (94.08%), EFI.vn (93.58%), SRMR.rel (88.80%), EFI (86.02%), TLI.rel (84.96%), RMSEA.rel (83.87%) and (CFI.rel 82.17%). The least accurate fit indices were, from best to worst: TLI (65.51%), CFI (57.45%), RMSEA (32.43%), and SRMR (27.31%). Figure 5 shows the distribution of the percentage correct per fit index.

![Figure 5. Percentage Correct per Fit Index. The horizontal lines indicate the median values, the numbers inside the plots are the mean values.](image-url)
Figure 6 shows how the accuracy varies per level of factor loadings and correlations. In general, the accuracy improves with the increase in factor loadings (overall accuracy of 54.41% for loadings of .40, 76.29% for loadings of .55 and 86.45% for loadings of .70) and decreases with higher factor correlation (overall accuracy of 87.41% for correlations of 0, 76.96% for correlations of .50 and 52.87% for correlations of .70).

TEFI.vn and EFI.vn presented an almost perfect accuracy for factor loadings of .55 and .70 ($PC_{TEFI.vn} = 98.41\%$, $PC_{EFI.vn} = 98.24\%$ and $PC_{TEFI.vn} = 99.96\%$, $PC_{EFI.vn} = 99.92\%$, respectively), and an impressive accuracy for factor loadings of .40 ($PC_{TEFI.vn} = 83.87\%$, $PC_{EFI.vn} = 82.56\%$). EFI presented a good accuracy for a loading of .40 (74.73%), and high accuracies for loadings of .55 and .70 (89.92% and 93.40%, respectively). TLI.rel, RMSEA.rel and CFI.rel presented lower accuracies for small loadings (68.13%, 62.78% and 59.86%, respectively), but high accuracies for loadings of .55 (93.44%, 92.38% and 90.79% respectively) and .70 (93.32%, 96.46% and 95.86% respectively).

Considering the size of the correlation, it is interesting to note that TEFI.vn, EFI.vn and TLI.rel presented a moderately high accuracy, even when the correlation is .70 ($PC_{TEFI.vn} = 87.61\%$, $PC_{EFI.vn} = 86.97\%$ and $PC_{TLI.rel} = 80.02\%$), followed by SRMR.rel (78.51%). EFI was the fifth best fit measure when correlations between factors was .70 (74.43%), followed by RMSEA.rel (70.82%) and CFI.rel (70.78%).

The skew of the items also plays a role in the accuracy. When the skew is -2 (skew 1 in Figure 6), the overall accuracy is 72.98% and when the skew is 2 (skew 4 in Figure 6) the overall accuracy is 75.30%. For the skew -2 and 2 (skew 2 in Figure 6), the overall accuracy was 66.61%, while for a skew ranging from -2 to 2, by increments of .5 (skew 3), the general accuracy was 74.63%. TEFI.vn and EFI.vn stands out once again in the worst condition in terms of skew, presenting an accuracy of 92.64% and 91.80% for the skew -2 and 2, respectively, followed by SRMR.rel (80.56%), RMSEA.rel (77.78%), CFI.rel (75.66%), TLI.rel (75.44%) and EFI (74.26%).

Overall, TLI (47.53%), CFI (49.44%), RMSEA (46.81%), and SRMR (44.56%) presented low accuracies, and will not be presented in detail. Their low accuracy can be explained by the values of these fit indexes, that has a tendency to be in a range considered adequate by the cut-off points used ($\geq .95$ for CFI and TLI; $\leq .05$ for RMSEA and SRMR). Figure 7 shows the distribution of the fit indices values across all combination of four items into two factor for factor loadings of .40 and correlation between factors of .70. The values of the traditional indices indicates that all the incorrect structures could be considered as presenting a good fit, if the cut-off values are used. Note that all tested structures present CFI and TLI values higher than .95 and RMSEA and SRMR values lower than .05. Notably, the CFI and TLI values are maximized when the correct structure is specified and when the RMSEA and SRMR values are minimized. This pattern helps to understand why using these indices as relative fit measures works much better than using their cut-off values.

**Shuffled, Underfactoring and Overfactoring Conditions**

In terms of the overall performance, the fit indices can be separated into four groups: (1) high accuracy (PC) in the four conditions tested (TEFI.vn), (2) high or moderately high PC in three of the four conditions (EFI, RMSEA.rel, CFI.rel, SRMR.rel, TLI.rel), (3) high PC in two conditions (EFI.vn), and (4) moderate or low PC across conditions.
Figure 6. Percentage correct per levels of correlation (columns), skew (rows), loadings (color), and fit index.
Figure 7. Distribution of the fit indices values across all combination of four items into two factors
(TLI, CFI, SRMR and RMSEA). TEFI.vn presented an overall accuracy of 92% (see Figure 8), followed by EFI (PC = 84%) and TLI.rel (PC = 92%). On the one hand, TEFI.vn presented very high PCs even in the most extreme overfactoring conditions (see Figure 9), especially with correlated factors (PC = 96.57%). On the other hand, TEFI.vn did not perform as well in the underfactoring condition with high factor loading and high interfactor correlation (PC = 58.95%). This problem does seem to vanish if the number of items is high (PC = 100%).

EFI presented a very high PC in the shuffled (99.68%) and underfactoring condition (99.98%), but only a moderately high PC in the overfactoring condition (85.72%) and a moderate PC in the more extreme overfactoring condition (52.45). EFI.vn, by its turn, presented a very high PC in the shuffled (100%) and underfactoring condition (100%), but basically zero accuracy in both overfactoring conditions. The CFA indices used as relative fit measures (RMSEA.rel, CFI.rel, TLI.rel and SRMR.rel) were stable across the shuffled, underfactoring and overfactoring conditions, presenting PCs ranging from 98% and 100%. However, in the more extreme overfactoring conditions their PCs dropped to 36.94% for TLI.rel, 26.23% for RMSEA.rel, 15.84% for CFI.rel and 0.50% for SRMR.rel. When applied using cut-off values only, the CFA fit measures presented low to moderate PCs in the shuffled, underfactoring and overfactoring conditions, and very low PC in the more extreme overfactoring condition (see Figure 8).

![Figure 8](image.png)

**Figure 8.** Mean percentage correct (PC) per fit index, in the overfactoring and underfactoring conditions.

Figure 9 shows a more detailed description of the performance of the fit indices across conditions. The accuracy of TEFI.vn is very high, except when the interfactor correlation is zero, or the correlation is moderate (.55) and the factor loading is .55, both in the extreme overfactoring condition. As mentioned before, the average accuracy of TEFI.vn is not high in the underfactoring condition, with high interfactor correlation and high loadings. Besides these three conditions, TEFI.vn presents a very
high accuracy to detect the simulated number of dimensions, even in the most extreme overfactoring condition, in which most of the methods present moderately low to low accuracy.

EFI, RMSEA.rel, CFI.rel, SRMR.rel and TLI.rel presented very high PCs across conditions when the loadings were .70, irrespective of the correlation between factors, condition (shuffled, underfactoring or overfactoring), sample size, and number of items. In the more extreme overfactoring condition and for factor loadings of .70, however, EFI was the most accurate in estimating the correct dimensionality structure than the other methods, presenting a PC of 62.95% for inter-factor correlations of 0, 69.75% for inter-factor correlations of 50, and 69.40% for inter-factor correlations of 70 (see Figure 9). Conversely, EFI presented PCs similar to the other CFA fit indices for factor loadings of .50, while EFI.vn failed almost in all overfactoring conditions.

In terms of the CFA fit indices, TLI.rel was the second most accurate fit index in the more extreme overfactoring condition. For factor loadings of .70 it presented a PC of 15.15% for inter-factor correlations of 0, 40.12% for inter-factor correlations of 50, and 50.31% for inter-factor correlations of 70.

**Applied Example:**

*The world health organization study on global ageing and adult health (SAGE)*

This empirical example investigates the dimensionality structure of a large dataset from the World Health Organization (WHO) Study on global Ageing and adult health (SAGE). Although being a longitudinal study focusing on adults aged 50 years and older, WHO’s SAGE also collects data from nationally representative adult samples.
Table 1. Fit Measures for the SAGE dataset

<table>
<thead>
<tr>
<th>Data</th>
<th>Model</th>
<th>Walktrap Steps</th>
<th>EFI</th>
<th>TEFI.vn</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA</th>
<th>SRMR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>EGA</td>
<td>3</td>
<td>-5.03</td>
<td>-69.51</td>
<td>0.92</td>
<td>0.92</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>EGAtmfg</td>
<td>6</td>
<td>-5.48</td>
<td>-52.73</td>
<td>0.92</td>
<td>0.91</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Theoretical Model</td>
<td></td>
<td>-4.89</td>
<td>-46.53</td>
<td>0.90</td>
<td>0.89</td>
<td>0.04</td>
<td>0.06</td>
</tr>
<tr>
<td>Test</td>
<td>EGA</td>
<td></td>
<td>-5.52</td>
<td>-68.61</td>
<td>0.92</td>
<td>0.92</td>
<td>0.03</td>
<td>0.06</td>
</tr>
<tr>
<td></td>
<td>EGAtmfg</td>
<td></td>
<td>-5.60</td>
<td>-52.70</td>
<td>0.92</td>
<td>0.91</td>
<td>0.03</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>Theoretical Model</td>
<td></td>
<td>-4.89</td>
<td>-46.01</td>
<td>0.90</td>
<td>0.89</td>
<td>0.04</td>
<td>0.06</td>
</tr>
</tbody>
</table>

(aged between 18 and 49 years) in China. In the current paper, only data from the first wave of data collection will be used.

The original sample is composed by 9,283 adults, aged between 18 and 99 years (Mean = 60.53, SD = 11.91), being 53.40% females. In terms of the highest level of education completed, 225.09% finished primary school, 18.22% high school (or equivalent), 27.96% secondary school, 21.82% less than primary school, and 6.7% have college/pre-university/university or post-graduate degree. The 9,283 participants answered all 52 questions related to broader domain labeled activities and participation, containing items related to six theoretical domains: (1) social, community and civic activities/involvement, (2) interpersonal relations, (3) mobility, (4) self-care, (5) domestic life and (6) personal history of work and employment.

The dataset was split into two random subsamples (training and testing set), each containing 50% of the data, for cross-validation. The results of both EGA estimations were compared to the theoretical dimensionality structure (six domains) briefly described above. The EFI, TEFI.vn and traditional factor-analytic fit measures were computed for each structure (i.e. the structures estimated using EGA with GGM, EGAtmfg, plus the theoretical model) in the training and testing sample. The TEFI.vn index was used to select the optimal number of steps in the walktrap algorithm (used in the EGA technique), comparing the dimensionality structure estimated via EGA using from three to 10 steps.

Table 1 shows the factor-analytic fit values of the theoretical model in the training (CFI = 0.90, TLI = 0.89, RMSEA = 0.04, SRMR = 0.06) and test set (CFI = 0.89, TLI = 0.89, RMSEA = 0.04, SRMR = 0.06), suggesting a poorer fit of this model compared with the other two empirical models.

The structure estimated via EGA presented the lowest TEFI.vn in both samples, compared with the theoretical structure and the structure estimated via EGAtmfg. Both EGA and EGAtmfg estimated five factors, but with a different composition of items per factor. The factor-analytic fit measures were basically identical for the factors estimated via EGA and EGAtmfg, while TEFI.vn was lower for the structure estimated using EGA in both samples (in the training and testing/validation sample).

Figure 10 shows the structure estimated via EGA and EGAtmfg, while Table 2 presents the variables, the EGA and EGAtmfg factor numbers per variable, the description of the variables and the theoretical factor per variable. Factor one (red nodes in Figure 10) represents a socialization factor, with items such as Q6001 (attending public meeting),
Q6005 (had friends over to your home) and Q6009 (went out of house to see friends or relatives), among others. Factor two (yellow/green-ish nodes) contains items related to work and employment (e.g., Q6020: Do/did you usually work throughout the year, or do/did you work seasonally, or only once in a while for your main job?) and to the participation in the political life (e.g., Q6021: Lots of people find it difficult to get out and vote. Did you vote in the last state/national/presidential election?). Factor three (light green nodes) contains items related to overall difficulty in the past 30 days with self-care (e.g., Q2036: In the last 30 days, how much difficulty did you have bathing/washing your whole body?), domestic work (e.g., Q2039: In the last 30 days, how much difficulty did you have day to day work?), mobility (e.g., Q2026: In the last 30 days, how much difficulty did you have walking 100 meters), and with engaging in social events (Q2033: In the last 30 days, how much difficulty did you have joining community activities?). Factor four (blue nodes) presents items related to overall difficulty in daily activities (work or household activities, moving around, doing vigorous activities, bathing, taking care of appearance and staying by yourself). Factor five, by its turn, contains items related to overall difficulty with interpersonal relations and activities.

![Figure 10. Structure of the SAGE data using EGA](image)

**Discussion**

Reducing the dimensions of multivariate data has proven to be an extremely worthwhile task in the study of behavior. This has generally taken the form of procedures inferring the number and nature of unmeasured (latent) variables (e.g., factors) that can account for the calculated relations among the observed variables. Willingness to ascribe causal properties to these latent variables varies between
investigators and is not the primary concern of this article; however, accurately estimating the number of dimensions (e.g., latent variables) is our primary concern.

For many researchers maintaining a causal latent variable perspective, the relations between latent variables can provide the generality needed to articulate useful lawfulness. An example used by Nesselroade and Molenaar (2016) concerned the frustration produces aggression hypothesis. Some students, frustrated because they cannot solve a difficult mathematics problem, might curse the teacher whereas other students might slam down their books and walk out of class and still others might slap at a fellow student nearby. To say that “inability to solve a difficult math problem leads to slapping a fellow student” might hold for some students but it does not have generality whereas at the latent variable level “frustration produces aggression” covers all three situations above. Obviously, the properties of frustration and aggression must be inferred from the observable variables such as cursing, walking out, and slapping, but to build lawfulness that has generality, latent variables are key. The relations among latent variables are central to constructing or inferring mechanisms that help build common explanations of behavior. Extracting the most out of one’s multivariate data relies heavily on model fitting that best estimates the dimensionality and nature of the latent variables. Therefore, powerful methods for fitting models to data and judging the goodness of that fit are indispensable.

It is common practice among applied researchers and methodologists to use traditional fit indices from factor analysis and structural equation modeling to assess the factor solutions of their data because it provides useful diagnostic information, but also enables the comparison of different dimensionality structures. As pointed out earlier, there are only a few simulation studies investigating the accuracy of traditional fit indices (such as CFI and RMSEA) to estimate the correct number of factors (Clark & Bowles, 2018; Frazier & Youngstrom, 2007; Garrido et al., 2016; Yang & Xia, 2015) and, in general, they suggest that the traditional fit indices should be used with caution, especially when analyzing dichotomous data. Clark and Bowles (2018), for example, concluded that the common fit statistics cut-offs are, at best, imperfect tools for guiding decisions regarding dimensionality. The final suggestion is to use fit indices in conjunction with other methods for assessing dimensionality. This suggestion is perhaps best exemplified by combining EGA with the EFIs to obtain optimal dimension identification.

The accuracy of the EFI fit measures in correctly identifying the number of simulated factors was compared to RMSEA, CFI, TLI, and SRMR in two Monte Carlo simulations. In the first simulation, two latent factors with four variables each were simulated and the fit indices were used to check the fit of 70 different combinations of variables within the two factors. The goal of the first simulation was to identify the accuracy of the fit indices in detecting the correct simulated structure, given that the number of factors is correct but the placement of items per factor isn’t. The results show that TEFI.vn and EFI.vn are the most accurate fit measures, followed by SRMR (when used as a relative measure of fit) and EFI. In general, the RMSEA/SRMR cut-offs presented very low accuracies in almost all conditions tested, except when the factor loadings were high and the interfactor correlation was zero or .50. Even in this favorable condition, the accuracy dropped for highly skewed items (skew 2; skew −2 and 2). For conditions with high interfactor correlation (.70), the accuracy of the traditional fit indices, using the cut-off values only, was very low. However, CFI, TLI, RMSEA and SRMR presented high accuracies, comparable to TEFI.vn, EFI.vn and
EFI, in almost all conditions when used as relative measures of fit.

In the second simulation, we investigated the accuracy of the fit indices to identify the correct structure compared to shuffled, underfactored and overfactored conditions. TEFI.vn presented a very high accuracy, in general, even in the most extreme overfactoring condition. The main pattern is that TEFI.vn works better with correlated factors, and improves its accuracy with higher factor loadings (except in the underfactoring condition with high correlation between factors). EFI presented a very high accuracy in the shuffled and underfactor conditions, and was the most accurate fit measure in the most extreme overfactoring condition (i.e. when one factor is split into two factors, without any items from other factors).

EFI.vn was also very accurate in the shuffled and underfactoring conditions, but completely failed in the overfactoring conditions. This result makes sense, since Von Neumann’s entropy was developed to understand entangled quantum states. In the current simulation, the underfactor condition can be thought as reflecting mixed “states”, since two simulated factors are merged as one single factor. Therefore, the EFI.vn value increases when two factors are merged into one, compared to the EFI.vn value of the original simulated structure. However, the EFI.vn measure could not detect the correct simulated structure when one factor is split into two factors. When a state is not mixed but instead separated, then EFI.vn value remains low because the states are still reflecting homogeneous states. This may also be a consequence of the strategy used to penalize the number of factors. Future studies should investigate different penalization methods to verify if it is possible to increase the accuracy of EFI.vn in the overfactoring conditions. A second explanation to the difference in accuracy between EFI.vn and the TEFI.vn is that the former replaced the joint entropy of the factors by the total entropy of the correlation matrix, which can be more robust in conditions where the partition of the multidimensional space is plausible but incorrect (i.e. extreme overfactoring condition).

The traditional fit indices, when used as absolute measures of fit, presented moderate to high accuracies in the shuffled and underfactoring conditions, with interfactor correlations of zero (orthogonal structures). Notably, the accuracies dropped with the increase of interfactor correlations, becoming very low for interfactor correlations of .70. A very similar pattern occurred for the overfactoring condition. However, when the overfactoring condition was more extreme (one factor split into two pure factors), the cut-off values presented low accuracies irrespective of the interfactor correlation and factor loadings. Finally, when used as relative measures, the traditional fit indices presented very high accuracies for all conditions, except for the more extreme overfactoring condition. These measures, which are usually thought of as absolute are actually not. Instead, our results show that they are better thought of as relative measures. Therefore, our results suggest that these measures should not be used as absolute but rather relative fit measures.

In sum, the current paper corroborates previous simulation studies, showing that the traditional factor-analytic fit indices should be avoided for dimensionality assessment purposes and not used as absolute measures of fit (based on cut-off values). They are, however, very accurate when used as relative measures of fit specifically for the shuffled, underfactoring, and non-extreme overfactoring conditions. Also, the three new fit indices (EFI, EFI.vn, and TEFI.vn) were accurate in identifying the simulated number of factors and the respective placement of items per factor in the shuffled and underfactoring conditions. Notably, only TEFI.vn presented a high accuracy in both
overfactoring conditions. These results represent an important advance in the area of measures of fit for structural and dimensionality analysis.

Finally, an empirical example was presented, using data from the WHO’s Study on Global Ageing and Adult Health (SAGE) from China. Exploratory graph analysis was used to estimate the dimensionality structure using the GGM and the TMFG approaches, and their fit were compared to the fit of a theoretical model, in two random subsamples. The $TEFI_{vn}$ index pointed to the EGA (GGM) structure (five factors) as the structure with the best fit to the data (compared to the theoretical model and to the structure estimated via EGAtmfg) in both samples (training and testing/validation). Basically, CFI, TLI, RMSEA and SRMR could not differentiate between the EGA and the EGAtmfg structures, both presenting five factors but with a different composition of items per factor.

Limitations

The limitations of the current paper can be summarized in two main aspects. The first one related to the data simulation approach used, which did not include correlated residuals (minor factors) at the population level. Moreover, only dichotomous data was used in the simulation. Future studies would bring new light into this area by incorporating other types of data (e.g., polytomous and mixed type, i.e., continuous, dichotomous and polytomous). Second, the entropy fit indices only work for multidimensional factors (i.e., structures with at least two factors) presenting a simple structure (one item linked only to one factor). EGA uses the walktrap algorithm, which is a technique that identify communities in weighted networks where one node (or variable/item) is part of only one community (or factor). The entropy fit indices were developed as a way to check the fit of different dimensionality structures estimated via EGA (and to compare them to theoretical structures). Future studies should extend the entropy fit measures to accommodate unidimensional structures and to allow cross-loadings. Moreover, future research should examine the potential to extend the EFIs to other structural equation modeling structures, which would greatly expand their utility to a larger number of researchers.
References


Garrido, L. E., Abad, F. J., & Ponsoda, V. (2016). Are fit indices really fit to estimate the number of factors with categorical variables? Some cautionary findings via


of the American Statistical Association, 58, 238–244.


Table 2.: Variables, factors estimated via EGA and EGAtmfg, description of the variables and the theoretical structure of the variables

<table>
<thead>
<tr>
<th>Item</th>
<th>EGA Factor</th>
<th>EGAtmfg Factor</th>
<th>Description</th>
<th>Theoretical Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>q6001</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you attended any public meeting in which there was discussion of local or school affairs?</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q6002</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you met personally with someone you consider to be a community leader?</td>
<td>Interpersonal</td>
</tr>
<tr>
<td>q6003</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you attended any group, club, society, union or organizational meeting?</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q6004</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you worked with other people in your neighborhood to fix or improve something?</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q6005</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you had friends over to your home?</td>
<td>Interpersonal</td>
</tr>
<tr>
<td>q6006</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you been in the home of someone who lives in a different neighbourhood than you do or had them in your home?</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q6007</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you socialized with coworkers outside of work?</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q6008</td>
<td>1</td>
<td>3</td>
<td>How often in the last 12 months have you attended religious services (not including weddings and funerals)?</td>
<td>Community, Social and Civic</td>
</tr>
</tbody>
</table>
q6009 1 3 How often in the last 12 months have you gotten out of the house/your dwelling to attend social meetings, activities, programs or events or to visit friends or relatives? Community, Social and Civic

q1508 2 5 Are/were you paid in cash or kind for your work or are/were you not paid at all? Work and Employment

q1509 2 5 Who is/was your employer in your current/most recent MAIN job? Work and Employment

q1511 2 5 Do/did you usually work throughout the year, or do/did you work seasonally, or only once in a while for your main job? Work and Employment

q6020 2 2 How interested would you say you are in politics and national affairs? Would you say you are .....? Community, Social and Civic

q6021 2 1 Lots of people find it difficult to get out and vote. Did you vote in the last state/national/presidential election? Community, Social and Civic

q6022 2 5 How much say do you have in getting the government to address issues that interest you? Community, Social and Civic

q6023 2 2 How free do you think you are to express yourself without fear of government reprisal? Community, Social and Civic

q2025 3 4 In the last 30 days, how much difficulty did you have sitting for long periods Mobility

q2026 3 4 In the last 30 days, how much difficulty did you have walking 100 meters Mobility

q2027 3 4 In the last 30 days, how much difficulty did you have standing up from sitting down Mobility
<table>
<thead>
<tr>
<th>q2028</th>
<th>3</th>
<th>4</th>
<th>In the last 30 days, how much difficulty did you have standing for long periods</th>
<th>Mobility</th>
</tr>
</thead>
<tbody>
<tr>
<td>q2029</td>
<td>3</td>
<td>4</td>
<td>In the last 30 days, how much difficulty did you have climbing flight of stairs</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2030</td>
<td>3</td>
<td>4</td>
<td>In the last 30 days, how much difficulty did you have stooping, kneeling, crouching</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2031</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have picking up things with fingers</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2032</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have taking care of household responsibilities</td>
<td>Domestic Life</td>
</tr>
<tr>
<td>q2033</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have joining community activities</td>
<td>Community, Social and Civic</td>
</tr>
<tr>
<td>q2034</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have extending arms above shoulder</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2036</td>
<td>3</td>
<td>4</td>
<td>In the last 30 days, how much difficulty did you have walking a long distance</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2037</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have bathing/washing your whole body</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2038</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have getting dressed</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2039</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have day to day work</td>
<td>Domestic Life</td>
</tr>
<tr>
<td>q2040</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have carrying things</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2041</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have moving around inside your home</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2042</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have with eating</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2043</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have with getting up from lying down</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2044</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have with getting to and using the toilet</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2045</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have with getting where you want to go, using transport</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2046</td>
<td>3</td>
<td>2</td>
<td>In the last 30 days, how much difficulty did you have getting out of your home</td>
<td>Mobility</td>
</tr>
<tr>
<td>q9006</td>
<td>3</td>
<td>3</td>
<td>Have difficulties walking?</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2001</td>
<td>4</td>
<td>1</td>
<td>Difficulty with work or household activities moving around</td>
<td>Domestic Life</td>
</tr>
<tr>
<td>q2002</td>
<td>4</td>
<td>1</td>
<td>difficulty in vigorous activities</td>
<td>Mobility</td>
</tr>
<tr>
<td>q2003</td>
<td>4</td>
<td>1</td>
<td>Self-care such as bathing</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2004</td>
<td>4</td>
<td>1</td>
<td>Grooming, taking care of appearance</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2005</td>
<td>4</td>
<td>1</td>
<td>Staying by yourself</td>
<td>Self-care</td>
</tr>
<tr>
<td>q2006</td>
<td>4</td>
<td>1</td>
<td>personal relationships/participation in community</td>
<td>Interpersonal</td>
</tr>
<tr>
<td>q2012</td>
<td>5</td>
<td>1</td>
<td>dealing with conflicts and tensions with others</td>
<td>Interpersonal</td>
</tr>
<tr>
<td>q2013</td>
<td>5</td>
<td>1</td>
<td>friendships</td>
<td>Interpersonal</td>
</tr>
<tr>
<td>q2014</td>
<td>5</td>
<td>1</td>
<td>dealing with strangers</td>
<td>Interpersonal</td>
</tr>
</tbody>
</table>