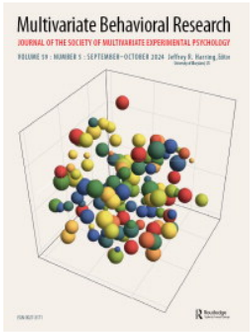

Multivariate Behavioral Research

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/hmbr20

Determining Sample Size Requirements in EFA Solutions: A Simple Empirical Proposal

Urbano Lorenzo-Seva & Pere J. Ferrando

To cite this article: Urbano Lorenzo-Seva & Pere J. Ferrando (2024) Determining Sample Size Requirements in EFA Solutions: A Simple Empirical Proposal, Multivariate Behavioral Research, 59:5, 899-912, DOI: 10.1080/00273171.2024.2342324

To link to this article: https://doi.org/10.1080/00273171.2024.2342324

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC

View supplementary material

Published online: 08 May 2024.

Submit your article to this journal

Article views: 4113

View related articles

View Crossmark data

Citing articles: 6 View citing articles

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=hmbr20

MULTIVARIATE BEHAVIORAL RESEARCH
2024, VOL. 59, NO. 5, 899–912
https://doi.org/10.1080/00273171.2024.2342324

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Determining Sample Size Requirements in EFA Solutions: A Simple Empirical Proposal

Urbano Lorenzo-Seva and Pere J. Ferrando

Deparment of Psychology, Universitat Rovira i Virgili, Tarragona, Spain

ABSTRACT

In unrestricted or exploratory factor analysis (EFA), there is a wide range of recommendations about the size samples should be to attain correct and stable solutions. In general, however, these recommendations are either rules of thumb or based on simulation results. As it is hard to establish the extent to which a particular data set suits the conditions used in a simulation study, the advice produced by simulation studies is not direct enough to be of practical use. Instead of trying to provide general and complex recommendations, in this article, we propose to estimate the sample size that is needed to analyze a data set at hand. The estimation takes into account the specified EFA model. The proposal is based on an intensive simulation process in which the sample correlation matrix is used as a basis for generating data sets from a pseudo-population in which the parent correlation holds exactly, and the criterion for determining the size required is a threshold that quantifies the closeness between the pseudo-population and the sample reproduced correlation matrices. The simulation results suggest that the proposal works well and that the determinants identified agree with those in the literature.

KEYWORDS

Sample size; exploratory factor analysis; unrestricted factor analysis

Introduction

Since the first applications of the unrestricted or exploratory factor analysis (EFA) model, the issue of the minimal sample size required to achieve a "proper" solution has been controversial and hotly debated (e.g. Cattell, 1952; Thurstone, 1940). No doubt, one of the main reasons for this controversy is the complexity of the problem. By way of illustration, consider first a simple, standard inferential statistical application based on a known reference distribution (e.g. estimating the population mean). In this setting, determining the sample size that is needed to achieve an unbiased and stable parameter estimate is generally straightforward. In contrast, consider now the EFA scenario. First of all, we shall define what is meant by a proper solution, which is, at least, a multifaceted concept. Thus, for example, we could focus on goodness of fit, and consider a proper solution to be one that provides correct fitting results, which, in turn, means correctly estimating the number of common factors. Alternatively, we can focus on replicability across studies (e.g. Cattell, 1952; Osborne & Fitzpatrick, 2012) or on the structural estimates obtained. In this latter case, we would consider a proper sample solution to be one that is close to the 'true' population. Even within this facet, however, several criteria of closeness can be considered. For example, the pattern congruence (Cattell, 1952; Comrey & Lee, 1992), or a measure of discrepancy between the 'true' and the observed pattern. Regarding this last criterion, it should be further noted that an EFA solution is not unique (i.e. the rotational indeterminacy problem), which makes any reference to the "true" pattern questionable.

Given the indeterminacy and the importance of the problem, it is no surprise that numerous recommendations, rules and/or cutoff values have been made regarding the minimum sample size (N) required for fitting an EFA solution since the model was first implemented. Simply put, two general classes of recommendations can be distinguished. First, there is a wide range of what are essentially rules of thumb.

CONTACT Urbano Lorenzo-Seva urbano.lorenzo@urv.cat Facultat de Ciències de l'Educació i Psicologia, Universitat Rovira i Virgili Ctra. de Valls s/n, Tarragona, Spain.

Supplemental data for this article can be accessed online at https://doi.org/10.1080/00273171.2024.2342324.

© 2024 The Author(s). Published with license by Taylor & Francis Group, LLC
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

These are generally reasonable but simplistic, and can be misleading in some cases. In this class, the most common rules are of two types: (a) overall or absolute, and (b) relative, based on the number of variables in the data set. In the first type, the usual recommendations range between $N = 200$ and $N = 1000$ (Comrey & Lee, 1992; Goretzko et al., 2021; Gorsuch, 1983; Mundfrom et al., 2005). In the second, the cases-per-variable ratios usually range between 4-to-1 and 20-to-1 (Bentler & Chou, 1987' Cattell, 1952; Tanaka, 1987).

While the cases-per-variable ratio can be considered an important determinant in terms of the stability and accuracy of an EFA solution, it does not act constantly but interacts with many other determinants (see below), so adherence to a fixed ratio of this type might be misleading (e.g. MacCallum et al., 1999). The overall rule, instead, makes more sense as a minimal basis, especially if we consider that fitting an EFA solution entails obtaining estimates based on estimates. More in detail, the process can be conceived as two-step (e.g. Muthén, 1993). In the first step, correlation estimates are obtained from the sample data. In the second step, structural estimates (mainly loadings and inter-factor correlations) are obtained based on the sample correlation matrix. It seems clear, then, that the stability of the factor solution depends on the stability of the correlations that serve as a basis (see e.g. Ferrando & Lorenzo-Seva, 2014; Loo, 1983), so structural estimates cannot be expected to be stable if the first-step correlation estimates are not. In this regard, then, a minimum sample size is a basic requirement if the sample correlations are to converge toward their population values and their standard errors are to be kept acceptably low. However, even at this first level, the issue is more complex than this, as the stability of the correlations also depends on a range of determinants such as the distribution of the variables, the presence of outliers, the homogeneity of the sample, and the magnitude of the correlations themselves (e.g. Schönbrodt & Perugini, 2013). Even more relevant here is the type of FA model fitted: the linear model based on product-moment correlations or the nonlinear model for ordered-categorical variables based on polychoric correlations. All other things being equal, far larger samples are required to obtain stable estimates in the polychoric case (see e.g. Ferrando & Lorenzo-Seva, 2014; Lorenzo-Seva & Ferrando, 2021a).

We turn now to the second class of recommendations. This class is more rigorous, and generally based on simulation studies that attempt to assess the role of sample size together with other variables that can potentially determine the stability of the solution (e.g. Browne & Cudeck, 1993; Hogarty et al., 2005; MacCallum et al., 1999, 2001). These studies have similar structures and arrive at similar results, which suggests that the main determinants of the stability of an EFA solution are now well known. As a summary, they are: sample size, ratio variables-per factor, amount of measurement error in the variables (or, reciprocally, amount of common variance), and strength and determinacy of the solution. This final determinant involves the number of indicators per factor, the magnitude of the loadings, and the extent to which the variables approach simple structure (which is related to the amount of rotational indeterminacy). Now, even when the determinants are known, the problem is that they act interactively and in a complex way, which means that simple recommendations derived from their results will not be useful for the practitioner (at least not directly).

General aims of the proposal and organization

The state of affairs summarized above is the basis for the present proposal. Rather than attempting to provide general (and necessarily complex) recommendations, our aim is to directly propose an empirical estimate of the minimal sample size needed for the data set that is to be factor analyzed. The proposal, which is described in detail below, (a) takes into account those characteristics of both the data set and the specified solution that are the most important in terms of stability and closeness to the 'true' solution, and (b) is fully empirical, and avoids theoretical developments at the cost of intensive simulation which, given the capabilities of modern computers, is now affordable. As far as we know, what we propose here is a new contribution.

The basic proposal has six steps: (1) The initial sample correlation matrix obtained from the real population is used to obtain an estimate of the population correlation matrix; (2) an FA solution is fitted to the matrix estimated in step (1) and the corresponding reproduced correlation matrix is obtained; (3) the correlation matrix estimated in (1) is used as a seed to produce a pseudo-population data set in which the estimated correlation matrix in (1) holds exactly; (4) samples of increasing size are drawn from the pseudo-population and, in each sample, the FA solution is fitted and the reproduced correlation matrix obtained; (5) the closeness between the reproduced correlation matrix in each sample (step 4) is

compared with the reference reproduced correlation matrix in step (2); and finally, (6) the required minimum sample size is set as that in which the discrepancy in the comparisons in (5) falls below a determined and pre-specified threshold. If the aim of the analysis is fully exploratory, only the number of observed variables is known and the number of factors of the FA solution fitted in step (2) must be determined automatically. If the aim is more confirmatory, the expected number of factors is then used to specify the solution in step (2). For the sake of clarity, we should point out that the terms "confirmatory" or "more confirmatory" are used here to refer to an unrestricted FA solution in which the number of common factors is specified in advance.

The approach summarized above is known as Simple Empiric NumEriCAl estimate of sample size in factor analysis (SENECA estimate).

The remainder of the article is organized as follows. First, the procedure will be described in detail. Second, it will be tested using three simulation studies.

Simple empiric numerical estimate of sample size in factor analysis (SENECA estimate)

Consider a set of m observed variables related to p common factors, the population standardized variance-covariance (i.e. correlation) matrix $\mathbf{\Sigma}$ $(m \times m)$ in the set of observed variables, and the corresponding estimate $\mathbf{R}$ $(m \times m)$ obtained in a sample of N observations from the score matrix $\mathbf{X}$ $(N \times m)$. When $\mathbf{R}$ is obtained from a large and representative sample of the population, it is expected to be a good estimate of $\mathbf{\Sigma}$.

The direct (canonical form) UFA model decomposes $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{\Gamma}' + \mathbf{\Psi}, \tag{1}$$

where $\mathbf{\Gamma}$ is the loading matrix $(m \times p)$, and $\mathbf{\Psi}$ is the diagonal matrix $(m \times m)$ with the unique variances on the main diagonal. When a UFA solution is fitted to sample data, the aim is to estimate the matrices $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ from the observed matrix $\mathbf{R}$. In terms of the sample estimate, matrix $\mathbf{R}$ is decomposed as (e.g. Browne & Cudeck, 1993).

$$\mathbf{R} = \mathbf{A}\mathbf{A}' + \mathbf{U} + \mathbf{E}, \tag{2}$$

where $\mathbf{A}$ $(m \times k)$ and $\mathbf{U}$ $(m \times m)$ are the corresponding estimates of $\mathbf{\Gamma}$ and $\mathbf{\Psi}$ in expression (1), and $\mathbf{E}$ $(m \times m)$ represents the amount of observed covariance in $\mathbf{R}$ that cannot be accounted for by the sample factor model (Browne & Cudeck, 1993; MacCallum et al., 2007).

Once a UFA solution has been fitted to an observed correlation matrix $\mathbf{R}$, the reproduced variance-covariance matrix is defined as,

$$\mathbf{R}^* = \mathbf{A}\mathbf{A}' + \mathbf{U}. \tag{3}$$

To assess the accuracy of how well the population factor model has been recovered in expression (3), the products $\mathbf{\Gamma}\mathbf{\Gamma}'$ and $\mathbf{A}\mathbf{A}'$ can be compared using the Root Mean Square of Residuals. Because this index is based on the correlation residuals, we shall denote it as CRMSR (e.g. Ogasawara, 2001). Assuming the factor model is correct, the larger sample $\mathbf{X}$ is, the closer to zero the value of CRMSR is expected to be, and it will only be zero when the sample is infinitely large (i.e. asymptotically). What the SENECA estimate aims to do is determine the size of the sample needed to achieve a particular threshold value for CRMSR. As different threshold values lead to different sample sizes, the researcher needs to decide the threshold value that will be used for a particular data analysis. So, note that the threshold parameter controls the degree of accuracy that is required in the solution.

Once the threshold value has been set, the SENECA sample size required to achieve this threshold is based on the following steps. In addition, Figure 1 summarizes the algorithm proposed in SENECA estimate.

Step 1: Obtain an initial small sample representative of the population

An initial sample $\mathbf{X}_i$ needs to be obtained and used to estimate the size recommended for the final sample. We propose that the minimum size for this initial sample $\mathbf{X}_i$ be $N_i = 100 + 2\times\ m$ (i.e. at least one hundred observations plus two observations for each measured variable) in the linear FA model. If variables are treated as ordered-categorical, and a nonlinear FA model is fitted, then we propose to use $N_i = 200 + 2\times$ m. This proposal of initial sample sizes is based on previous pilot tests, and also on our experience of analyzing real data sets.

Next, let $\mathbf{R}_\mathrm{i}$ be the correlation matrix among the columns in $\mathbf{X}_i$. As the initial sample is small, the correlation coefficients r_{ij} in $\mathbf{R}_i$ are not expected to properly estimate the true correlation values in $\mathbf{\Sigma}$. To adopt a conservative position, the 95% confidence interval is computed for each r_{ij}, and the lower threshold value is taken as the initial estimate of the corresponding value in $\mathbf{\Sigma}$. The estimated values are

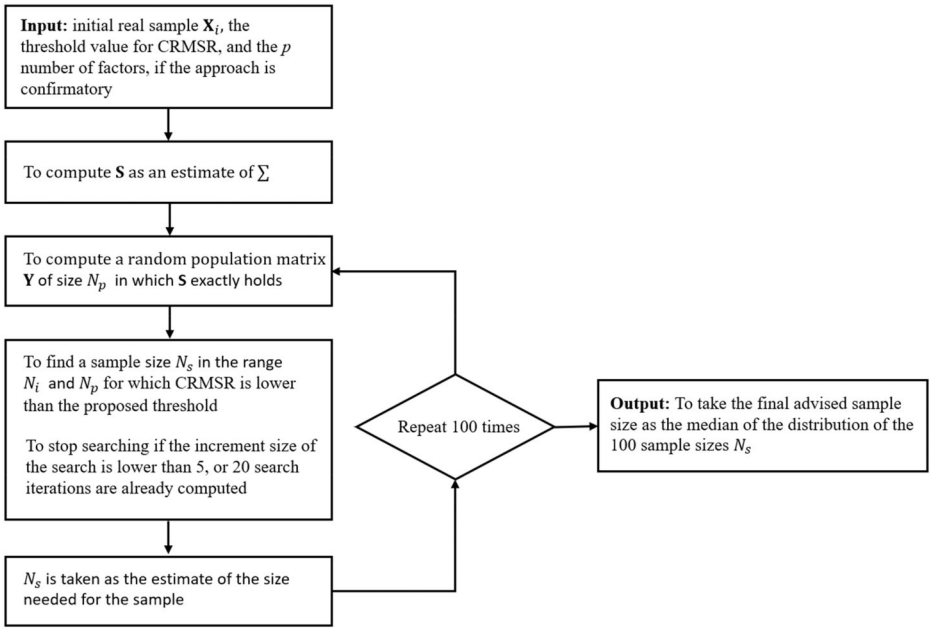

Figure 1. Flow char of the algorithm proposed in SENECA estimate.

collected in an **S** matrix that is expected to be a conservative estimate of **Σ**. If **S** turns out to be nonpositive definite, it can be smoothed using some of the procedures available (see Lorenzo-Seva & Ferrando, 2021a).

Step 2: Compute a population matrix Y of size N_p in which S exactly holds

Let's assume that matrix **S** obtained in the previous step is a true population matrix. If an EFA solution is fitted to **S**, then a population factor solution is obtained as in expression (1), and the product $\mathbf{\Gamma\Gamma}'$ represents the common variance in **S** (i.e. the best reproduced common variance that a sample can obtain).

If **Z** ($N_p \times m$) is a random matrix with normally distributed columns $N(0, 1)$ and **L** is Cholesky's decomposition of **S**, the product

$$\mathbf{V} = \mathbf{ZL} \tag{4}$$

produces a pseudo-population data matrix $\mathbf{V}(N_p \times m)$ in which the model (1) described above exactly holds. From this pseudo-population matrix **V**, random samples of observed scores can be drawn for which the corresponding correlation matrix is an estimate of the true population matrix **S**. Cholesky's method is a decomposition of a positive-definite matrix into the product of a lower triangular matrix and its conjugate transpose, and is frequently used in Monte Carlo simulations to produce random population samples based on a proposed correlation matrix. A description of this procedure can be found, for example, in Burgess (2022).

As a consequence of expression (4), the distribution of the variables in **V** are continuous and normally distributed. If the original variables in **X** are continuous, matrix **V** becomes **Y** without any further transformation. In the ordered-categorical model, the thresholds used to discretize the continuous variables generated in **V** are the same thresholds that were estimated in the parent sample data. So, the discrete variables generated have the same distributional characteristics as the original distributions in $\mathbf{X}_i$. After these transformations, matrix **V** in (4) becomes matrix **Y**. Then **R**, the correlation matrix between columns of **Y**, can be factor analyzed using expression (3).

The number of factors extracted from **R** depends on the aim of the analysis. If the aim is more confirmatory and a number of p factors are expected, then these p factors are the ones to be retained. If the aim is fully exploratory and the number of common factors cannot be specified in advance, then we propose that Kaiser's rule be used (i.e. to extract as many factors as eigenvalues larger than one in **S**). While this rule is known to overestimate the number of common factors, this is not undesirable here, since most of the common variance contained in the data needs to be considered. So it can be used here as a conservative bound.

To assess how accurately the pseudo-population factor model has been recovered, the products $\mathbf{\Gamma\Gamma}'$ (related to **S**) and $\mathbf{AA}'$ (related to **Y**) are compared using the correlation root mean squared residual

(CRMSR) index. As the CRMSR value in the pseudo-population is expected to be closer to zero than the threshold value proposed in a particular analysis, the size of N_p has to be large enough for this to be so (e.g. $N_p = 100{,}000$). If the value of CRMSR is larger than the threshold value, then the value of N_p must be increased, and the second step repeated until the value observed for the CRMSR is equal to or lower than the proposed threshold value.

Step 3: Find a sample size N_s with a value between N_i and N_p

The aim now is to determine the sample size that allows the observed CRMSR to be equal to or lower than the threshold value proposed. The value of N_s must be between N_i and N_p. The following fast algorithm can be used to find the optimal value:

a. Make $N_a = N_i$, $N_b = N_p$, $a = N_b - N_a$, $inc = a/4$, and $N_s = N_i$.
b. Make $N_s = N_a$, to obtain a subsample of the first N_s observations from $\mathbf{Y}$, fit the FA solution to obtain $\mathbf{A}_i$, and compute the corresponding CRMSR between products $\mathbf{\Gamma\Gamma}'$ (related to S) and $\mathbf{A}_i\mathbf{A}_i'$. If the CRMSR is larger than the proposed threshold, then, increase $N_s = N_s + inc$, and repeat this step.
c. Make $N_b = N_s$, and $a = a/2$. If $N_b - a < N_i$, then $N_a = N_i$ and $a = N_b - N_a$; otherwise $N_a = N_b - a$.
d. Make $inc = a/4$. If the algorithm has iterated fewer than 20 times or inc is larger than 5, then go to step b.

After 20 iterations or when the value of $inc < 5$, the sample size sought is the value of N_s. The number of 20 iterations is arbitrary and could be higher if N_p is very large. The more iterations there are, the more accurate the final value of N_s will be.

Step 4: Achieve a stable value for the final recommended sample size

As the value of N_s is obtained from a single population matrix $\mathbf{Y}$ in step 3, it is probably too risky to decide that the sample size that is really needed is the actual value of N_s. To obtain a more stable estimate, steps 2 and 3 should be repeated a number of times. In the analyses carried out here, we replicated steps 2 and 3 one hundred times. To decide on the final sample size, we propose using the median of the distribution of N_s values obtained in the replications of steps 2 and 3.

First simulation study

The aim of the first simulation study was to assess the performance of SENECA when different numbers of both observed variables and factors are involved in the model. The level of communality was also considered in the study.

A Monte Carlo simulation study was carried out using samples drawn from a true population model. Based on expression (1), a population loading matrix was defined. The parameters that were manipulated in the simulation study were:

1. Number of factors (p): between 1 and 5. When the number of factors of the loading matrix at hand was larger than 1, the inter-factor correlation values were uniform randomly drawn from the range [–.3; .3].
2. Number of observed variables (m): for each factor in the loading matrix to be generated, the number of variables in each factor were uniform randomly drawn from the range [5; 30]. This means that the number of variables does not perfectly correlate with the number of factors. For example, one loading matrix could be defined by 30 variables and a single factor, while another could be defined by 25 variables and 5 factors.
3. Salient loadings: for each loading matrix, a range of salient loadings was specifically defined. The lowest and highest salient standardized loadings were uniform randomly drawn from the interval [.35; .90]. Once the range had been defined, salient loadings for each variable were uniform randomly chosen from this range. Standardized nonsalient loadings were uniform randomly drawn from the interval [–.05; .05].
4. Nonsalient variables: for half of the loading matrices, 5% of the observed variables of the pattern at hand were set to be nonsalient for any factor. All the standardized loading values for these variables were uniform randomly drawn from the interval [–.05; .05].

For each population loading matrix, we checked that no Heywood cases were present and that the reproduced population correlation matrix was positive definite. From each true population matrix $\mathbf{\Lambda}$, 1,000 pseudo-population matrices were obtained with $N = 50{,}000$. In the first step, a correlation matrix $\mathbf{R}^*$

was obtained as $\mathbf{R}^{*} = \mathbf{\Lambda}\ \mathbf{\Phi}\mathbf{\Lambda}' + \mathbf{\Theta}^2$, where $\mathbf{\Lambda}$ is the population loading matrix, $\mathbf{\Phi}$ is the population inter-factors correlation, and $\mathbf{\Theta}$ is a diagonal matrix with unicities. Then we computed the Cholesky decomposition of $\mathbf{R}^{*}=\mathbf{L'L}$, where $\mathbf{L}$ is an upper triangular matrix. The pseudo-population data matrix of continuous variables $\mathbf{X}$ was finally obtained as $\mathbf{X} = \mathbf{ZL}$, where $\mathbf{Z}$ is a matrix of random standard normal scores, with rows equal to the corresponding sample size and a number of columns equal to the corresponding number of variables.

The first $100 + 2\times\ m$ observations of the pseudo-population matrix were taken as the initial sample and the SENECA estimate was computed with the exploratory approach (i.e. the number of factors was not specified), and the confirmatory approach (i.e. the true number of factors was specified). In addition, five levels of accuracy were required for the threshold value of CRMSR: 0.05, 0.04, 0.03, 0.02, and .01. Appendix shows the pseudo-code in R language that was used to obtain the simulated of data for some given population matrices $\mathbf{\Lambda}$ and $\mathbf{\Phi}$.

Once the SENECA estimate had been made, the initial sample of $100 + 2 \times m$ observations was increased to achieve to the number of observations suggested by the SENECA estimate for both (exploratory and confirmatory) approaches, and for each threshold value of CRMSR. These two (exploratory and confirmatory) samples were factor analyzed and the RMSRs between the population loading matrix and the two sample loading matrices were computed.

The SENECA estimate was used to produce a total of 5 (number of factors) × 5 (number of threshold value of CRMSR) × 2 (exploratory and confirmatory approach) × 1,000 (replicas)= 50,000 samples. We implemented the simulation study in MATLAB R2023a MathWorks Inc (2007).

Results of the first simulation study

We observed that the communality level played an important role in the size the SENECA estimate indicated the samples should be. When the communality level was low, the sizes were much larger, and other characteristics of the data set (i.e. number of observed variables and number of factors in the population model) did not have much influence on the outcome. This was particularly the case in the exploratory approach. However, when the communality level was high, the more complex the model was, the larger the sample size had to be. To visualize this outcome, we present the outcomes in terms of low and high communality situations. We used a communality of 0.60 as the threshold between low and high communality levels.

Table 1. Descriptive statistics on the recommended sample sizes and accuracy of the sample loading matrix of solutions with communalities larger than 0.60.

CRMSR	*p*	Exploratory approach					Confirmatory approach				
		N		RMSR			*N*		RMSR		
		Mean	*SD*	Mean	*SD*	RMSR ≤ CRMSR, %	Mean	*SD*	Mean	*SD*	RMSR ≤ CRMSR, %
.05	1	140	17	.029	.010	97	137	14	.030	.010	96
	2	212	38	.035	.007	99	189	14	.043	.036	95
	3	254	33	.034	.004	100	239	16	.042	.051	96
	4	287	26	.033	.010	99	273	13	.044	.052	92
	5	309	21	.035	.014	98	298	14	.045	.044	90
.04	1	168	37	.027	.008	92	155	18	.028	.009	89
	2	324	76	.028	.008	99	281	29	.035	.032	93
	3	396	53	.028	.005	99	365	24	.041	.067	92
	4	445	44	.028	.012	98	419	17	.040	.065	93
	5	474	41	.029	.011	96	451	15	.039	.044	89
.03	1	278	83	.022	.006	91	253	52	.023	.007	87
	2	570	148	.022	.010	97	486	52	.026	.020	94
	3	709	108	.021	.003	98	647	44	.036	.078	91
	4	780	84	.023	.017	96	736	34	.032	.053	92
	5	840	79	.024	.019	94	798	30	.040	.068	85
.02	1	611	215	.015	.004	90	546	119	.015	.005	87
	2	1351	475	.015	.004	95	1087	141	.021	.033	91
	3	1643	342	.015	.004	93	1436	90	.044	.137	87
	4	1795	274	.017	.011	92	1644	74	.033	.087	87
	5	1920	254	.017	.007	84	1780	63	.036	.058	76
.01	1	2441	754	.007	.002	91	2223	471	.007	.002	89
	2	5348	1869	.010	.008	73	4379	487	.012	.015	64
	3	6704	1753	.012	.022	67	5747	363	.026	.078	63
	4	7472	1807	.012	.006	61	6578	306	.028	.073	60
	5	7838	1369	.014	.016	52	7086	245	.035	.077	50

Table 1 shows statistics for high levels of communality. It shows the mean, standard deviation sample size recommended by the SENECA estimate in terms of the level of accuracy required, and the number of factors of the population loading matrix. As can be observed, the exploratory approach recommended larger samples than the confirmatory approach (i.e. number of common factors specified in advance). The reason is that the second approach adds more information, so samples can be smaller. The number of factors also seems to play an important role: the larger the number of factors, the larger the sample needs to be.

In addition to the results above, the more demanding the required accuracy is, the larger the sample size needs to be (see e.g. Schönbrodt & Perugini, 2013). When the accuracy level is CRMSR = 0.05, the recommended sample sizes are quite small, but when CRMSR = 0.01 they are much larger. It should also be noted that when CRMSR is 0.02 or lower some of the recommended sample sizes are extremely large.

When the sample loading matrix and the population loading matrix (i.e. the values of RMSR) were used and the CRMSR threshold was equal to or larger than 0.03, the accuracy was even higher than required. When the CRMSR threshold was equal to 0.02 and the approach was exploratory, the accuracy achieved and the accuracy required were of about the same order. Finally, when the CRMSR threshold was equal to 0.01, the accuracy achieved was slightly worse than the accuracy required. The confirmatory approach was observed to behave in a similar way only when the value of CRMSR was larger than .03. For other levels of CRMSR, only when the number of factors was low than the accuracy of RMSR was lower than the CRMSR.

We were also interested in assessing when the RMSR was lower than CRMSR. If RMSR was systematically equal to or lower than CRMSR, then, our strategy of setting a value for CRMSR with the hope that it will control the RMSR will be reinforced. With this aim, we inspected the percentage of times that RMSR was lower than CRMSR. The table shows that in the exploratory approach, the percentage was very high (above 0.90) when CRMSR was larger than .01. However, when CRMSR was 0.01, it was only acceptable when the number of factors was 1. In the case of the confirmatory approach, the outcome was not so positive: only when CRMSR was larger than 0.03 was the percentage larger than .89. When CRMSR was 0.03, the outcomes were acceptable except if the number of factors was 1 or 5. For other levels of CRMSR, the percentage could turn out to be too low (e.g. when CRMSR was 0.01 and $p = 5$). Our conclusion is that the exploratory approach is preferable when the accuracy is the most important aspect of the analysis. When the confirmatory approach is used, samples might be smaller but the accuracy achieved may be worse.

Table 2. Descriptive statistics on the recommended sample sizes and accuracy of the sample loading matrix of solutions with communalities lower than 0.60.

CRMSR	*p*	Exploratory approach					Confirmatory approach				
		N		RMSR			*N*		RMSR		
		Mean	*SD*	Mean	*SD*	RMSR ≤ CRMSR, %	Mean	*SD*	Mean	*SD*	RMSR ≤ CRMSR, %
.05	1	219	79	.036	.011	91	144	10	.045	.013	73
	2	371	96	.034	.009	99	211	13	.049	.032	82
	3	376	77	.035	.004	99	248	12	.049	.035	89
	4	374	58	.036	.005	98	273	10	.049	.034	90
	5	374	46	.038	.011	98	293	15	.047	.025	89
.04	1	347	279	.030	.009	88	201	18	.037	.010	68
	2	612	174	.027	.008	98	321	20	.040	.027	79
	3	611	144	.028	.005	97	380	17	.040	.032	88
	4	599	113	.029	.005	97	414	14	.044	.044	88
	5	604	89	.030	.009	97	437	13	.042	.029	86
.03	1	617	371	.022	.007	88	350	34	.028	.007	70
	2	1206	399	.021	.009	95	565	38	.033	.033	77
	3	1193	325	.022	.006	95	669	32	.036	.051	85
	4	1159	264	.024	.013	93	729	27	.043	.064	78
	5	1147	207	.024	.012	92	768	25	.038	.042	78
.02	1	1395	906	.015	.005	85	782	75	.018	.005	67
	2	2905	1191	.016	.016	93	1258	86	.026	.039	71
	3	3031	1062	.016	.012	93	1494	70	.033	.079	76
	4	2948	792	.017	.015	86	1626	59	.035	.063	74
	5	2907	670	.019	.014	83	1715	60	.038	.058	70
.01	1	5326	2670	.008	.003	85	3098	327	.009	.002	64
	2	11610	4417	.010	.012	79	5028	341	.016	.026	42
	3	12489	4730	.012	.014	67	5962	295	.030	.092	29
	4	13465	4686	.012	.013	60	6461	247	.040	.097	21
	5	13250	4046	.014	.014	54	6817	228	.035	.069	21

Table 2 shows statistics for low levels of communality. The exploratory approach now systematically asks for larger samples, and the size of the sample turns out to be very large in some cases (e.g. when CMRSR is 0.01 and the number of factors is equal to 5). In addition, the sample sizes recommended for a particular value of CRMSR did not seem to be related to the number of factors, except if the number of factors was 1 (in which case the sample size recommended was substantially lower). However, the accuracy levels achieved were similar to those when the communality level was high. This outcome seems to suggest that large sample sizes were recommended because this was the only way to achieve the accuracy required.

In the confirmatory approach, the sample sizes recommended were similar to those when communality was high. It could also be observed that the higher the number of factors, the larger the sample sizes recommended. However, accuracy and the percentage of times that RMSR was equal to or lower than CRMSR were only acceptable when CRMSR was 0.05.

Table 3 shows the correlation matrix between the recommended sample sizes and some of the variables that characterize the population loading matrix. In the exploratory approach, the recommended sample size correlated positively with the number of factors and number of observed variables, and negatively with the communality level. However, there was a difference between low and high communality levels. When the communality was low, the number of variables and the number of factors were the most decisive characteristics. Likewise, the greater the accuracy required, the less important the number of variables and factors were, and the more important the communality was. But when the communality was low, the communality was the most decisive characteristic.

In general, a similar pattern of correlations was observed when the number of factors was specified (i.e. the more confirmatory approach). In this case, however, the correlation involving the communality level is generally close to zero and the number of factors in the population shows the largest correlation at all levels of accuracy. It must be said, however, that in the simulation study the true number of factors was known. In real situations, the more confirmatory approach could be problematic if the specification of the factors does not correspond with the true number of factors in the population.

A total of 5% of variables were forced to be nonsalient in half of the population loading matrices. As pointed out above, the number of observed variables is related to the recommended sample size. To assess the presence of nonsalient variables (regardless of the number of observed variables), we computed the partial correlation between the number of nonsalient variables and the sample size recommended by Seneca Estimate, controlling for the actual number of observed variables. The outcomes are shown in Table 4. It turned out that the number of nonsalient variables only has a negative effect when the approach is confirmatory and the communality is low: the higher the number of nonsalient variables, the larger the recommended sample size.

Our overall conclusion from the simulation study is that the sample sizes recommended by the SENECA estimate are appropriate for recovering the population loading matrix with a degree of accuracy equal to or better than that required a priori, except when requirements are very demanding (i.e. CRMSR = 0.01). We therefore recommend a sensible accuracy level of CRMSR = 0.03.

To illustrate how the three hyperparameters used in SENECA (i.e. maximum iterations, *inc*, and initial sample size) interact between them, we identified a situation in our simulation study in which the advice of SENECA failed to produce the expected precision, and inspected it properly. It was a situation of 5 factors in the population, 123 observed variables and average communality of .56. The initial 346 observations of our pseudo-population were used to estimate

Table 3. Correlations between sample sizes recommended by SENECA estimate and characteristics of the population loading matrix.

h^2	CRMSR	Exploratory approach		Confirmatory approach			
		m	*p*	h^2	*m*	*p*	h^2
>.60	.05	.823	.889	−0.132	.909	.962	−0.037
	.04	.741	.851	−0.209	.832	.949	−0.105
	.03	.725	.828	−0.135	.832	.942	−0.044
	.02	.645	.751	−0.254	.833	.939	−0.127
	.01	.637	.723	−0.205	.844	.940	−0.084
<.60	.05	.364	.452	−0.505	.855	.943	−0.001
	.04	.276	.355	−0.447	.801	.929	.016
	.03	.305	.373	−0.488	.798	.925	.014
	.02	.339	.390	−0.488	.799	.926	.013
	.01	.460	.488	−0.488	.795	.922	−0.009

Table 4. Correlations between sample sizes recommended by SENECA estimate and the number of nonsalient variables in the population loading matrix after controlling for number of observed variables.

CRMSR	Exploratory approach		Confirmatory approach	
	$h^2 < .60$	$h^2 > .60$	$h^2 < .60$	$h^2 > .60$
.05	.102	.032	.403	.044
.04	.087	.052	.431	.078
.03	.121	.055	.441	.053
.02	.122	.055	.525	.119
.01	.165	.034	.549	.053

the final sample size using SENECA, with a CRMSR = 0.01 (without specifying the number of factors expected). SENECA needed 12 iterations to advise to use a sample of 7,250 observations. However, when we used the first 7,250 observations of our pseudo-population to compute the corresponding factor analysis, the RMSR between the sample loading matrix and the population loading matrix was of 0.0325 (i.e. the error observed was larger than the pretended with our CRMSR threshold). We run again SENECA with the same initial 346 observations, now with the hyperparameters maximum iterations = 100 and $inc < 2$. SENECA needed 14 iterations to advise to use a sample of 7,205 observations. Again, when we used the first 7,250 observations of our pseudo-population to compute the corresponding factor analysis, the precision obtained between sample and population loading matrices was RMSR = 0.035.

We suspected that our initial observations in the pseudo-population were not very much representative of the whole pseudo-population: it can be the case because it was a random sample from the pseudo-population. To inspect if this was the case, we randomized the observations in the pseudo-population and used an initial sample of 4,000 observations. We run again SENECA, now with the hyperparameters maximum iterations = 20 and $inc < 5$. SENECA needed 12 iterations to advise to use a sample of 7,018 observations. Now, when we used the first 7,018 observations of our pseudo-population to compute the corresponding factor analysis, the precision obtained between sample and population loading matrices was RMSR = 0.009 (i.e. the expected for the precision level). Our conclusion was that the performance of SENECA depends on the representativeness of the initial sample, especially when a large accuracy is aimed.

Overall, the results discussed agree with those obtained in the simulation studies referred to above (e.g. Browne & Cudeck, 1993; Hogarty et al., 2005; MacCallum et al., 1999, 2001). As mentioned, the larger the communality is, and the more the solution approach's simple structure (i.e. it is determinate), the smaller the sample needs to be to attain an accurate and stable solution. Also, our results agree with previous simulations in that the interactions between the determinants can be quite complex. It is for this reason that no general or practical rules for determining sample size were provided by the previous studies. Instead, the advantage of our proposal is that it recommends a sample size which is directly based on the characteristics of the data set at hand. And, in addition, the required degree of accuracy of the solution can be modulated to some extent by the user.

Second simulation study

The aim of the second simulation study was to assess the performance of SENECA across different values of the elements in the inter-factor correlation matrix. We replicated a specific condition ($p = 3$, CRMSR = 0.03) from the first simulation study. However, for this second simulation study, the values in the inter-factor correlation matrix were defined as follows:

1. Zero correlation: off-diagonal values in $\mathbf{\Phi}$ were set to zero.
2. Low correlation: off-diagonal values in $\mathbf{\Phi}$ were uniformly drawn from the range [.10; .30].
3. High correlation: off-diagonal values in $\mathbf{\Phi}$ were uniformly drawn from the range [.40; .60].

As in the previous study, the number of replications of the study were 1,000. The SENECA estimate was used to produce a total of 3 (levels of inter-factor correlations) × 2 (exploratory and confirmatory approach) × 1,000 (replicas) = 6,000 samples.

Results of the second simulation study

Table 5 shows the mean, standard deviation sample size recommended by the SENECA estimate in terms of the level of accuracy required, and the number of factors of the population loading matrix. If the inter-factor correlation matrix does not play a role in the

Table 5. Descriptive statistics on the recommended sample sizes and accuracy of the sample loading matrix of solutions when inter-factor correlations are manipulated.

h^2	PHI	Exploratory approach					Confirmatory approach				
		N		RMSR		RMSR <= CRMSR, %	*N*		RMSR		RMSR <= CRMSR, %
		Mean	*SD*	mean	*SD*		Mean	*SD*	mean	*SD*	
>.60	0	708	121	.021	.001	98	643	42	.032	.070	94
	.10−.30	702	116	.022	.014	97	642	42	.029	.057	95
	.40−.60	708	121	.024	.013	96	642	44	.046	.109	89
<.60	0	1191	319	.022	.014	97	667	33	.034	.044	86
	.10−.30	1194	319	.022	.001	96	666	30	.037	.054	86
	.40−.60	1185	319	.022	.001	97	669	32	.035	.045	83

sample sizes that SENECA recommends, then the values observed in Table 5 should be similar to the corresponding conditions in Tables 1 and 2. In the case of the exploratory approach, the sample sizes advised for the corresponding conditions had a mean of 709 (when communality was large) and 1,193 (when the communality was low). The outcomes obtained in this second study shown similar sample sizes. The values of the statistics that assess precision are also comparable.

In the case of the confirmatory approach, the sample sizes advised for the corresponding conditions had a mean of 647 (when communality was large) and 669 (when the communality was low). The outcomes obtained in this second study also shown similar recommended sample sizes. However, a difference was observed int he precision statistics: when the correlation between factors was large, the precision was slightly decreased. The worse precision levels were obtained when the communality level was low.

As a conclusion, large inter-factor correlation values do not seem to affect the performance of SENECA when the approach is exploratory. However, when a confirmatory approach is used, and the inter-factor correlation values are large, the sample size advised by SENECA might not be large enough to recover the loading matrix with the desired precision. The second conclusion is that, when the number of factors is known and, in addition, these factors are expected to be strongly correlated among them, then it would be advisable to compute SENECA based on an exploratory approach.

Third simulation study

The aim of the third simulation study was to assess the performance of SENECA when the initial sample was not perfectly representative from the population. Again, we replicated a specific condition ($p = 3$, CRMSR = 0.03) from the first simulation study. However, for this third study we included a condition in which the initial sample was not perfectly representative:

1. Initial sample representative: in this condition, the study replicates exactly the same procedure that was used in the first study.
2. Initial sample not perfectly representative: in this condition, the third of the observations from the initial sample were substituted by a sample from an alternative population in which a single factor solution was correct (i.e. $p = 1$). In the alternative population, the number of observed variables and the communality level of the variables was preserved.

If the representativeness of the initial sample does not play an important role in the sample size advised by SENECA, then the sample sizes advised in both situations should be similar. The number of replications was 1,000.

Results of the third simulation study

The average of sample sizes advised by SENECA when the initial samples were representative of the population were 1,044 (std = 363) and 955 (std = 251) for the exploratory and confirmatory approaches, respectively. When the initial sample was not representative, the average of the sample size advised for the exploratory approach was 955 (std = 251), which means that the error of approximation in the population, already lead to a substantial decrease of the sample size. In this scenario, SENECA advised a sample size that was actually lower than the one need for the characteristics of the population. In the case of the confirmatory approach, a substantial decrease of the sample was also observed: the mean of the advised samples was 649 (std = 34).

The conclusion of the third simulation study is that researchers must try strongly to compute SENECA based on an initial sample that is chosen to be as representative from the population as possible.

Implementation

We implemented the SENECA estimate in R (R script "SenecaEstimate.r"). This script uses only native functions in R, so no packages need to be downloaded. To use it, researchers have to store participants' responses in a text file, update the name of the input file, and execute the script. The parameters to be configured are (1) the number of replications in the simulation study, (2) the number of expected factors (if the aim of the analysis is confirmatory), and (3) the threshold value for the CRMSR.

In addition, we implemented SENECA in our software to compute factor analysis (Ferrando & Lorenzo-Seva, 2017), which can be downloaded free from the site https://psico.fcep.urv.cat/utilitats/factor/. This is offered as an extension analysis that can be computed before a specific EFA solution is fitted. In this software, the item response format can be treated as continuous-unbounded (linear model) or ordered-categorical

(nonlinear model), and the parameters that can be configured are the same as in the R code.

Discussion

Somewhat simplistically, sample size recommendations for fitting EFA solutions are either rules of thumb or based on complex simulation results. In our experience as advisors, none of these recommendations completely fulfill the need of the applied researcher who is undertaking a factor analytic study. The rules of thumb are too simplistic and misleading, while the simulation results are too complex to lead to practical advice. As we see it, what practitioners need is an initial estimate of the sample size that they will need if they are to obtain an appropriate solution. This is the type of advice our proposal aims to provide.

The main idea of the proposal is to use the data set under study as a seed to generate a pseudo-population data set in which a factor solution holds exactly. Then, samples are generated and the sample size required for the 'true' and the sample reproduced correlation matrices to be closer (in terms of discrepancy) up to a pre-determined threshold value is determined. While this basic idea is simple, the proposal has a certain technical complexity and requires intensive simulation. It must be said that the amount of computing time is not prohibitive in the data set in which the SENECA estimate was tested.

It should be stressed that the aim of the SENECA estimate is only to produce an indicative sample size, so that researchers can get an idea of the magnitude of the sample needed to reliably compute a factor analysis. Also, researchers have to take a variety of decisions that can affect the magnitude of the size recommended. In principle, the larger the initial sample is, the more accurately the population correlations will be estimated, and the more realistic the sample size recommendation will be. In addition, if a particular number of factors is expected (i.e. a more confirmatory approach), the sample size needed is more likely to be smaller. When the fully exploratory approach is used, the characteristics of the data analyzed (e.g. a large number of variables with low communality) may lead to very large samples being recommended. The pre-determined level of accuracy that is required also affects the recommended sample size. In our analysis, we concluded that (a) a threshold value of CRMSR = 0.05 can be too lenient; (b) a threshold value of CRMSR = 0.01 can be too demanding; and (c) a threshold value of CRMSR = 0.03 is likely to be appropriate in most applications. However, researchers must decide if, for a particular data set, other values should be specified. For example, if both the number of variables and expected factors is low, and the communality high, then the threshold could be set to be more demanding (e.g. 0.02 or 0.01). At the other extreme, if the number of variables and expected factors is large, and the communality low, then the threshold can be relaxed (e.g. 0.04 or 0.05) to obtain a sample of a reasonable size (while bearing in mind that the reproduced correlation matrix may not to be estimated very accurately). Alternatively, in this complex situation, researchers could also try to identify the noisiest variables and discard them from the analysis (Ferrando et al., 2023). If this is done correctly, the likely outcome will be a "cleaned" data set consisting of fewer variables with higher communalities.

In its most basic form (Eqs. (1)–(4)), the method we propose is intended for a direct or canonical solution, and the reproduced correlation matrix is obtained as the product of the canonical pattern matrix and its transpose. In this case, the CRMSR as a measure of accuracy of the preproduced correlation matrix directly quantifies also the accuracy that the canonical loading estimates will attain. In a correlated -factors solution, however, the link between the accuracy of the reproduced matrix and that of the structural estimates is not so direct, and is a topic that should be addressed in the future. Possibly, in this case, the CRMSR as a measure of distance should be complemented by an index of profile pattern similarity (i.e. congruence), and the inter-factor correlation matrix should also be taken into account. Furthermore, developing more elaborated threshold measures should also take into account the intended use of the EFA application. If the model is only used as a rough precursor of a more restricted solution, then, the congruence or pattern similarity would be, possibly, the most relevant indicator for setting threshold values as it was made in the simulation studies. In this case, the relations are not so direct, and it would be indeed of interest to establish a link between the accuracy at the reproduced correlation level and the accuracy in the structural parameters of the solution. However, this is a difficult issue, and we prefer to leave it for future research. In this respect, we also believe that such a combined index should probably take into account not only measures of distance (i.e. discrepancy) but also of pattern profile similarity (i.e. congruence). Conversely, if the unrestricted solution is of interest "per se" (and not as a mere precursor), then, obtaining accurate structural

estimates is as important as it can be in a restricted solution.

If the number of common factors can be plausibly expected in advance, SENECA estimate can be computed based on the confirmatory approach. However, if the expected factors have been observed to be highly correlated in previous studies, then the results of our simulation study suggest that it could be more adequate to compute SENECA based on the exploratory approach. It must also be pointed out that, to compute SENECA estimate based on the confirmatory approach to decide the sample size, and then to compute a pure exploratory factor analysis (e.g. using procedures aimed at assessing the dimensionality of the data set) seems like and uncoherent analysis schema. If a pure exploratory factor analysis is to be computed, then, SENECA should also be computed based on the exploratory approach.

A further potential limitation of the threshold approach discussed so far is that it is based on the sample CRMSR, which is known to overestimate its population parameter when both the sample size and the degree of misfit of the proposed solution decrease (Maydeu-Olivares, 2017). So, in the future if would be of interest to assess the performance of modified versions of the CRMSR that have been found to produce less biased estimates (Maydeu-Olivares, 2017).

Computing SENECA estimate from the expected population number of factors is a good way of obtaining the optimal sample size when the control of the precision level is not an important feature of the study at hand. However, if the number of population factors is not available, then it is preferable to consider a large number of factors in the sample model so that the level of communality is not underestimated. As the extra factors in the sample model are not truly substantial factors in the population, they will not add much to the communality, and the estimate sample size will increase less than if a low number of factors are considered in the sample. In SENECA estimate, we opted to base our method on Kaiser's rule precisely because it has a tendency to retain more factors than the true number in the population, and hardly ever less. Once the final sample is available, then using a more reliable method (e.g. parallel analysis) to determine the number of factors is advisable. Another point of view is that the extra cases usually suggested in the exploratory approach seem to help the loading matrix have the desired level of precision.

When a large sample is recommended by SENECA estimate for a particular data set at a given precision level, researchers should consider if a lower precision is acceptable for their analysis. They should also determine whether the set of variables meet the minimum quality needed to be safely factor analyzed. For example, if a subset of observed variables shares a low amount of communality with the overall set of observed variables, then this subset of variables should not be included in the analysis (see e.g. Ferrando et al., 2023; Lorenzo-Seva & Ferrando, 2021b; for a discussion of this topic).

Our algorithm requires to decide the number of iterations and the amplitude of the interval in which to search for the final advised sample size. We tested our algorithm with specific values (20 maximum iterations and $inc < 5$). However, the researcher is free to compute SENECA based on more demanding values (e.g. 1,000 maximum iterations and $inc < 2$). These settings might increase the accuracy of the sample size advised by SENECA, but might also demand a large amount of computing time. In addition, the outcome of our third simulation study suggests, that the accuracy provided by SENECA depends on the representativeness of the initial sample that is used.

SENECA incorporates an algorithm to find an optimal sample size once minimum and maximum sample sizes have been identified. However, other search algorithms such as the approach based on surrogate modeling could be applied. It would be interesting to assess if other search approaches can improve the computing speed of Seneca Estimate.

As correlation estimates based on polychoric correlations are less stable, the sample size suggested by Seneca will be larger than the estimates based on Pearson correlations. Further research is required on how the number of response categories and the distribution of responses (e.g. skewness and kurtosis) can affect the sample size recommended by Seneca.

Any methodological proposal runs the risk of being misused, and this is particularly so in this case. Note that it is the sample data set under analysis that is used to generate the pseudo-population and run the procedure. Indeed, if the sample is inappropriate and not representative of the population for which the analysis is intended (e.g. a convenience sample, a range-restricted sample, or a mixture), then using what is proposed here is doomed to provide invalid or misleading results. So, it is the researchers' responsibility to use a proper design and collect a representative sample. The same considerations apply to the final sample used in the study. If the design and the sample are both appropriate, it is suggested that the tool proposed here is useful for factor-analytic applications.

Article information

Conflict of Interest Disclosures: Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

Ethical Principles: The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

Funding: This work was supported by Grant PID2020-112894GB-I00 from the MICIN/AEI/10.13039/501100011033.

Role of the Funders/Sponsors: None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

Acknowledgments: The authors would like to thank Albert Maydeu-Olivares for his comments on prior versions of this manuscript. The ideas and opinions expressed herein are those of the authors alone, and endorsement by the authors' institutions or the funding agency is not intended and should not be inferred.

References

Bentler, P. M., & Chou, C. P. (1987). Practical issues in structural modeling. *Sociological Methods & Research*, *16*(1), 78–117. https://doi.org/10.1177/0049124187016001004

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Sage.

Burgess, N. (2022). *Correlated Monte Carlo simulation using Cholesky decomposition.* SSRN.

Cattell, R. B. (1952). Factor Analysis, Harper and Row, New York, NY.

Comrey, A., & Lee, H. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Earlbaum Associates Publishers.

Ferrando, P. J., & Lorenzo-Seva, U. (2014). El análisis factorial exploratorio de los ítems: Algunas consideraciones adicionales. *Anales de Psicología*, *30*(3), 1170–1175. https://doi.org/10.6018/analesps.30.3.199991

Ferrando, P. J., & Lorenzo-Seva, U. (2017). Program FACTOR at 10: Origins, development and future directions. *Psicothema*, *29*(2), 236–240. https://doi.org/10.7334/psicothema2016.304

Ferrando, P. J., Lorenzo-Seva, U., & Bargalló-Escrivà, M. T. (2023). Gulliksen's pool: A quick tool for preliminary detection of problematic items in item factor analysis. *PLoS One*, *18*(8), e0290611. https://doi.org/10.1371/journal.pone.0290611

Goretzko, D., Pham, T. T. H., & Bühner, M. (2021). Exploratory factor analysis: Current use, methodological developments and recommendations for good practice. *Current Psychology*, *40*(7), 3510–3521. https://doi.org/10.1007/s12144-019-00300-2

Gorsuch, R. L. (1983). *Factor analysis* (2nd ed.) Lawrence Erlbaum Associates, Inc.

Hogarty, K. Y., Hines, C. V., Kromrey, J. D., Ferron, J. M., & Mumford, K. R. (2005). The quality of factor solutions in exploratory factor analysis: The influence of sample size, communality, and overdetermination. *Educational and Psychological Measurement*, *65*(2), 202–226. https://doi.org/10.1177/0013164404267287

Loo, R. (1983). Caveat on sample sizes in factor analysis. *Perceptual and Motor Skills*, *56*(2), 371–374. https://doi.org/10.2466/pms.1983.56.2.371

Lorenzo-Seva, U., & Ferrando, P. J. (2021a). Not positive definite correlation matrices in exploratory item factor analysis: Causes, consequences and a proposed solution. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(1), 138–147. https://doi.org/10.1080/10705511.2020.1735393

Lorenzo-Seva, U., & Ferrando, P. J. (2021b). MSA: The forgotten index for identifying inappropriate items before computing exploratory item factor analysis. *Methodology*, *17*(4), 296–306. https://doi.org/10.5964/meth.7185

MacCallum, R. C., Browne, M. W., & Cai, L. (2007). Factor analysis models as approximations. In R. Cudeck & R. C. MacCallum (Eds.), *Factor Analysis at 100: Historical developments and future directions* (pp. 167–190). Routledge. https://doi.org/10.4324/9780203936764

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36*(4), 611–637. https://doi.org/10.1207/s15327906mbr3604_06

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

MathWorks Inc (2007). MATLAB - The Language of Technical Computing, Version 7.5. The MathWorks, Inc., Natick, Massachusetts. http://www.mathworks.com/products/matlab/.

Maydeu-Olivares, A. (2017). Assessing the size of model misfit in structural equation models. *Psychometrika*, *82*(3), 533–558. https://doi.org/10.1007/s11336-016-9552-7

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

Muthén, B. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen and J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Sage Publications.

Ogasawara, H. (2001). Standard errors of fit indices using residuals in structural equation modeling. *Psychometrika*, *66*(3), 421–436. https://doi.org/10.1007/BF02294443

Osborne, J. W., & Fitzpatrick, D. C. (2012). Replication analysis in exploratory factor analysis: What it is and why it makes your analysis better. *Practical Assessment, Research, and Evaluation*, *17*(1), 15.

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Tanaka, J. S. (1987). How big is big enough?: Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*(1), 134–146. https://doi.org/10.2307/1130296

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin*, *37*(4), 189–236. https://doi.org/10.1037/h0059402

Appendix

Pseudo-code in R language to obtain the simulated of data for some given population matrices Λ and Φ

```
# L is the loading matrix in the population
# PHI is the inter-factor correlation matrix
# in the population
p <- ncol(L)
m <- nfil(L)
# To compute RP: the correlation matrix in
# the population (RP)
Rr <- L %*% PHI %*% t(L)
m <- ncol(Rr)
RP <- Rr - diag(diag(Rr)) +
diag(diag((matrix(1,m,m))))
# To compute Xsp: the pseudo-population
# matrix of m observed variables
# and 50,000 rows
L <- chol(RP)
Z <- matrix(0,50000,m)
for (i in 1:m) Z[,i]=rnorm(50000,0,1)
Xsp <- Z %*% L
# Nmin: the minimum sample size
Nmin <- 100+m*2
# To obtain the first Nnim rows from the
# pseudo-population matrix of m
# observed variables (Xsp)
X <- as.matrix(Xsp[1:Nmin,1:m])
# To compute SENECA based on X with a CRMSR =
# 0.03 with an exploratory
# approach to obtain the advised sample size
# (Nexp)
(Nexp) <-SenecaExploratory(X,0.03)
# To obtain the first Nexp rows from the
# pseudo-population and
# to compute the observed correlation matrix
Xexp <- as.matrix(Xsp[1:Nexp,1:m])
Rexp <- cor(Xexp)
# To compute SENECA based on X with a CRMSR =
# 0.03 with a confirmatory
# approach to obtain the advised sample size
# (Nconf)
Nconf <-SenecaConfirmatory(X,p,0.03)
# To obtain the first Nconf rows from the
# pseudo-population
# and to compute the observed correlation
# matrix
Xconf <- as.matrix(Xsp[1:Nconf,1:m])
Rconf <- cor(Xconf)
```