Commentary

Data reduction analyses of animal behaviour: avoiding Kaiser's criterion and adopting more robust automated methods

F. Blake Morton[a*], Drew Altschul[b,d,e]

[a] Psychology, School of Life Sciences, University of Hull, Hull, U.K.

[b] Scottish Primate Research Group, U.K.

[c] Department of Psychology, The University of Edinburgh, Edinburgh, U.K.

[d] Centre for Cognitive Ageing and Cognitive Epidemiology, Edinburgh, U.K.

Received 30 May 2018

Initial acceptance 14 August 2018

Final acceptance 14 December 2018

MS number 18-00356

***Correspondence:** F. B. Morton, Psychology, School of Life Sciences, University of Hull, Hull HU67RX, U.K.

E-mail address: b.morton@hull.ac.uk

Data reduction analyses such as principal components and exploratory factor analyses identify relationships within a set of potentially correlated variables, and cluster correlated variables into a smaller overall quantity of groupings. Because of their relative objectivity, these analyses are popular throughout the animal literature to study a wide variety of topics. Numerous authors have highlighted 'best practice' guidelines for component/factor 'extraction', i.e. determining

how many components/factors to extract from a data reduction analysis, because this can greatly impact the interpretation, comparability and replicability of one's results. Statisticians agree that Kaiser's criterion, i.e. extracting components/factors with eigenvectors >1.0, should never be used, yet, within the animal literature, a considerable number of authors still use it, even as recently as 2018 and across a wide range of taxa (e.g. insects, birds, fish, mammals) and topics (e.g. personality, cognition, health, morphology, reproduction). It is therefore clear that further awareness is needed to target the animal sciences to ensure that results optimize structural stability and, thus, comparability and reproducibility. In this commentary, we first clarify the distinction between principal components and exploratory factor analyses in terms of analysing simple versus complex structures, and how this relates to component/factor extraction. Second, we highlight empirical evidence from simulation studies to explain why certain extraction methods are more reliable than others, including why automated methods are better, and why Kaiser's criterion is inappropriate and should therefore never be used. Third, we provide recommendations on what to do if multiple automated extraction methods 'disagree' which can arise when dealing with complex structures. Finally, we explain how to perform and interpret more robust and automated extraction tests using R.

Key words: factor analysis, Kaiser's criterion, parallel analysis, principal components analysis, scree plot

Data reduction analyses such as principal components analysis (PCA) and exploratory factor analysis (EFA) identify relationships within a set of potentially correlated variables, and cluster correlated variables into fewer groupings called ‘components’ (in PCA) or ‘factors’ (in EFA) (Gorsuch, 1983; Field, 2009). Because they provide researchers with a relatively objective approach to categorizing different sets of data (e.g. questionnaire ratings, task performances or rates of behaviour among individuals), such analyses are commonly used to study a wide variety of theoretical and applied topics on animals (e.g. genetics, health, sociality, personality and cognition).

Numerous authors within the statistical literature have highlighted ‘best practice’ guidelines for component/factor ‘extraction’, i.e. determining how many components/factors should be extracted from a data reduction analysis, because this can greatly impact the interpretation, comparability and replicability of structures derived from these analyses (e.g. Zwick, & Velicer, 1986, Todorov, Fournier, & Gerber, 2018). Most notably, statisticians largely agree that one extraction method, Kaiser’s criterion, should never be used because it increases the risk of overextraction compared to more automated tests, which in turn can lead to instability in the structures derived from data reduction analyses, and thus affect the overall interpretation of one’s results. In terms of animal research, for example, Stevens, De Groot, and Staes (2015) subjected bonobo, *Pan paniscus*, social relationship data to a data reduction analysis and compared structures derived using Kaiser’s criterion versus a more robust and automated method called parallel analysis (discussed below in further detail). These authors found that the latter approach led to a more stable and conservative structure (two rather than three components), thereby changing the interpretation of their results entirely.

There are multiple extraction methods, mostly but not exclusively quantitative, that researchers can use as more robust alternatives to using Kaiser’s criterion to identify the quantity of underlying latent variables, i.e. those factors that are not directly observed but can be inferred from the data. That said, throughout the animal literature a considerable number of authors still use Kaiser's criterion to extract components/factors despite decades of resolve within the statistical literature, which is probably fuelled by the fact that it remains the 'default' method in common statistical packages such as SPSS (Field, 2009). Studies using Kaiser's criterion have been published as recently as 2018, encompassing an eclectic range of taxa, such as insects, birds, fish, and mammals, and covering a broad range of topics, including but not limited to personality (e.g. Martin & Reale, 2008; Menzies, Timonin, McGuire, & Willis, 2013; Pritchard, Sheeran, Gabriel, Li, & Wagner, 2014; Slipogor, Gunhold-de Oliveira, Tadic, Massen, & Bugnyar, 2016), cognition (e.g. Keagy, Savard, & Borgia, 2011; Meulman & van Schaik, 2013), morphology (e.g. Yakubu & Okunsebor, 2011; Dunham, Maitner, Razafindratsima, Simmons, & Roy, 2013; Khargharia, Kadirvel, Humar, Doley, Bharti, & Das, 2015), behavioural ecology (e.g. Adamo, Kovalko, & Mosher, 2013; Hassrick, Crocker, & Costa, 2013; Nath, Singha, Deb, Das, & Lahkar, 2015; Willems, Arseneau, Schleuning, & van Schaik, 2015; Klein, Pasquaretta, Barron, Devaud, & Lihoreau, 2017), sociality (e.g. Fraser, Schino, & Aureli 2008; Schino, & Aureli, 2008; Fraser & Bugnyar, 2010; McFarland & Majolo, 2011; Rebecchini, Schaffner, & Aureli, 2011; Fraser, Koski, De Vries, Van de Kraats, & Sterck, 2012; Moreno, Highfill, & Kuczaj, 2017;), welfare (e.g. Ferreira, Mendl, Guilherme, et al., 2016), health and conservation (e.g. Morton, Todd, Lee, & Masi, 2013; de Medeiros Filho, de Carvalho-Neto, Garcia, et al., 2018), reproduction (e.g. Venturini, Savegnago, Nunes, et al., 2013), life history (e.g. Poinapen, Konopka, Umoh, et al., 2017), acoustics and communication (Finger, Bastian, & Jacobs, 2017) and inbreeding (e.g. Lawrence, Mastromonaco, Goodrowe, et al., 2017). It is therefore clear that further awareness is needed to ensure that researchers of animal behaviour are reporting results that optimize structural stability and, thus, comparability and reproducibility of those results by making careful decisions about component/factor extraction.

In this commentary, we first clarify the distinction between principal components and exploratory factor analyses in terms of analysing simple versus complex structures, and how this relates to component/factor extraction. Second, we highlight recent empirical evidence from simulation studies to explain why certain extraction methods are more reliable than others, including why automated methods are better, and why Kaiser's criterion is inappropriate and should never be used. Third, we provide recommendations on what to do if multiple automated extraction methods 'disagree' which can arise when dealing with complex structures. Finally, we explain how to perform and interpret more robust and automated extraction tests in R.

<H1> PCA or EFA, Simple or complex structure?

Deciding which extraction methods are appropriate in a data reduction analysis depends on whether PCA or EFA is used, and whether the underlying structure of one's solution is simple versus complex. PCA and EFA are often applied interchangeably, but the theoretical foundations of the two methods are different. For instance, PCA attempts to account for the total variance (Velicer, 1976), but unlike PCA, EFA does not assume that variables have been measured without error (Brown, 2009). PCA is also a pure data reduction technique, which generates parsimonious summary variables that are linear combinations of the observed variables (Velicer, 1976). As there is no theory associated with this approach, there is technically no 'true' number of components that a researcher can extract. On the other hand, EFA is premised on having a theoretical model or models, in which latent variables cause the observed variables. This type of analysis fits a model using the correlation matrix of the observed data to account for common variance, i.e. the variance in a variable that is shared with other variables (Costello & Osbourne, 2005). These are just a handful of many differences between PCA and EFA, and so for interested readers, we recommend Brown (2009) and Yong and Pearce (2013) for beginners, and Gorsuch (1983) and Velicer and Jackson (1990) for more experienced researchers.

Historically, researchers have used PCA and EFA interchangeably for data reduction in animal behaviour research without issue because the results are very often the same. However, there is no guarantee of this, and if researchers wish to search for meaningful latent variables, then EFA should be used, and methods for identifying a meaningful number of factors should also be used (Fabrigar, Wegener, MacCallum, & Strahan, 1999). In the context of some studies, like those examining social relationship structure, the goal has been to identify underlying latent variables, which implies that researchers are theoretically justified in using EFA. As such, PCA should generally not be used. For this reason, we refer only to factors throughout this commentary, although when earlier works have used PCA, we refer to their results in terms of components. For a comparable guide to the use of PCA, we recommend Todorov et al. (2018).

If a researcher posits a theoretical structure to their data, a question they must also ask themselves is whether this structural model is simple or complex. A simple model is one in which variables tend to load strongly on one factor and weakly on all others (Revelle & Rocklin, 1979). Simple structure also implies that the model has only one ‘level’. More complex models, i.e. those that contain more than one level, include hierarchical models in which one or more higher-order factors are loaded on by lower-order factors, or bifactor models, in which a parallel factor is loaded on by the variables independently of the main lower-order factors (Murray & Johnson, 2011). For comparative examples of these models in animal behaviour and cognition, we recommend Arden and Adams (2016). If a researcher's theoretical model does not have a single level structure, EFA should not be used and the researcher should consider using, for example, confirmatory factor analysis (CFA) or a structural equation modelling (SEM) framework; we return to CFA and SEM in a subsequent section.

EFA assumes a single level structure, but it does not assume simple structure. If the researcher wishes to maximize the possibility of simple structure, usually because simple structure is easier to interpret, they could do this by allowing factors to correlate. This can be accomplished by specifying what is called an 'oblique rotation'. Rotations refer to the relationships between factors in space; the alternative to an oblique rotation is an orthogonal rotation. Factors that are orthogonal in space, e.g. x- and y-axes, have zero correlation (Jolliffe, 1986). However, there is rarely a theoretical reason for factors to have zero correlation in animal behaviour research and these factors are unlikely to have simple structure. Thus, if researchers are unsure or do not have justification, then an oblique rotation should be used (Browne, 2001).

<H1>Pros and cons of different extraction methods

As we have mentioned, a critical decision one must make before completing a data reduction analysis is how many factors to extract. This choice will influence how variables cluster together, thereby affecting the final solution and, hence, researchers' interpretation of those results (Zwick & Velicer, 1986; Ledesma & Valero-Mora, 2007). Underextraction can result in the loss of relevant information and distort the overall solution (Zwick & Velicer, 1986). Overextraction can result in some factors being unstable, making the overall solution difficult to interpret and/or replicate (Zwick & Velicer, 1986).

Deciding when to stop extracting factors depends on several competing considerations. As we have briefly touched on, and describe more fully below, there is a suite of quantitative and qualitative tools available to assist researchers in making this decision. However, researchers must also consider theory in EFA and look to the interpretability of the factors they extract. Even if all quantitative indicators suggest that a certain number of factors would yield the best model, the pattern of loadings between the latent and observed variables must be interpretable and the model should be theoretically viable. In other words, if variables representing distinct constructs load on a single factor, and/or variables representing the same construct load across many different factors, then the model will be theoretically uninterpretable and of little use (Fabrigar et al., 1999).

<H2>Kaiser's criterion

Various cutoffs have been developed to help researchers choose their factors, which typically involve taking into consideration the amount of variation that is explained by each factor (called 'eigenvalues'). As previously discussed, one problematic method that is still commonly used throughout the animal literature is Kaiser's criterion, which retains components with eigenvalues >1.0, that is, components/factors that account for more variance than what is accounted for by one of the original variables (Kaiser, 1960). Compared to other extraction methods, Kaiser's criterion is only appropriate to use with components, not factors, although researchers are not always aware of this nuance and have used Kaiser's criterion with EFAs (Costello & Osbourne, 2005). Moreover, unlike other techniques, Kaiser's criterion is largely arbitrary: there is little empirical reason why a component with an eigenvalue slightly greater than 1 ought to be retained while a component with an eigenvalue just below 1 should not

(Courtney, 2013). A component with an eigenvalue less than 1 accounts for less variance than the average observed variable, which is a reasonable criterion for exclusion, but it is too crude. Kaiser's criterion has shown tendencies towards overextraction and, to a lesser degree, underextraction (Zwick & Velicer, 1986). These biases are in part due to the observation that the number of components retained by the criterion reflects the number of variables included in the analysis more strongly than any attributes of underlying latent variables (Gorsuch, 1983). Ruscio and Roche (2012) simulated data from abstract theoretical models with varying numbers of factors and, for each simulation, tested several methods to determine how often each method selected the 'correct' number of factors as defined by the theoretical models. In these simulations, Kaiser's criterion led to a success rate of 8.77% and failed to extract the correct number of factors in more than 90% of cases (Ruscio & Roche, 2012).

Structures with high loadings (i.e. |0.7|) and/or those with components/factors containing four or more loadings greater than |0.4| are typically considered robust and reproducible (e.g. Guadagnoli & Velicer, 1988), yet studies relying on Kaiser's criterion do not always find this, which may be due to overextraction. Thus, simply put, no study should be using Kaiser' criterion to analyse their data.

<H2>Cattell's scree test

Another commonly used extraction method is Cattell's scree test, which is a graphical technique that plots eigenvalues in a simple line plot. The number of factors to extract is visually estimated from the scree plot by finding the point where the line drops and begins to level off; all components to the right of this point are considered random 'noise' and should therefore be excluded (Cattell, 1966). Within the animal literature, scree tests are often used alongside

Kaiser's criterion because, like Kaiser's criterion, they are the 'default' method in common statistical packages such as SPSS (Field, 2009).

Although scree tests are relatively simple to implement (perhaps contributing to their common usage by researchers), they are fundamentally subjective and, as such, can lead to spurious solutions. When factors are simple, observed variables load highly on one factor and there are few cross-loadings. Therefore, scree plots work well in such cases, as shown in Fig. 1a, because the solution is clearly discernible. On the other hand, when factors become more complex, scree plots open researchers to the risk of under- or overextraction due to their subjectivity, particularly as the line of the plot begins to asymptote, as shown in Fig. 1b (Zwick & Velicer, 1986).

In simulations, scree tests are correct in only 41.7% of cases (Zwick & Velicer, 1986). Thus, researchers should avoid using scree tests by themselves or alongside Kaiser's criterion, and only use them alongside more automated methods as a 'tie-breaker' if the plot reveals a distinct and unambiguous drop in eigenvalues past a certain component/factor (discussed in further detail below).

<H2>Automated extraction methods

Many alternative extraction methods have been developed that are more robust and automatic than Kaiser's and scree tests, and we strongly urge animal researchers to use them for data reduction analyses. Popular ones include the empirical Bayesian information criteria or empirical BIC (Schwarz, 1978), standardized root mean square residuals or SRMR (Hu & Bentler, 1999), Revelle and Rocklin's (1979) very simple structure (VSS) and Horn's (1965) parallel analysis (PA).

Empirical BIC is an information theoretical assessment of fit that evaluates the parsimony of any model (Schwarz, 1978). A solution with more components/factors will very often have a better absolute fit, but the BIC applies a penalty based on the number of parameters. Therefore, models with the lowest BIC are preferred. Because solutions with more components/factors have more parameters, BIC measures are an effective statistic for comparing many models. BIC is widely used in model building across different fields and is a superior statistic among information theory measures (Posada, Buckley, & Thorne, 2004). In simulations, BIC identifies the correct number of factors more than 60% of the time (Ruscio & Roche, 2012).

SRMR is the square root of the difference between a sample's covariance matrix and the proposed model's covariance matrix (Hooper, Coughlan, & Mullen, 2008). SRMR is representative of measures typically used in confirmatory factor analysis and is biased towards overextraction; however, the greater the number of parameters in the model and the larger the sample size, the lower SRMR tends to be (Hu & Bentler, 1999). Lower values are better; any value above 0.1 is considered unacceptable. To the best of our knowledge, SRMR has not been compared to alternative modern methods in simulation studies (Courtney, 2013).

VSS examines how well the individual components/factors fit within many solutions, where each progressive solution has one more factor than the last (Revelle & Rocklin, 1979). VSS can be used in an entirely objective fashion, by finding maxima, but it can be viewed subjectively as well, like a scree plot. However, VSS is best at identifying simple structures (i.e. those with a single level of factors) and therefore it is probably not appropriate if the 'true' structure of the data includes more than two factors (Revelle, 2015). To the best of our knowledge, VSS has not been compared to alternative modern methods in simulation studies (Courtney, 2013).

PA is based on generating random eigenvalues that 'parallel' the observed data in terms of sample size and the number of variables (Zwick & Velicer, 1986). A component/factor is retained if its eigenvalue is greater than the 95th percentile of the distribution of eigenvalues generated from the random data (Horn, 1965). This technique improves upon most other methods, both subjective (e.g. scree test) and objective (e.g. empirical BIC, Complexity), by taking sampling error into account, which is not partitioned from total variance in other methods (Horn, 1965). PA is not arbitrary: the 'parallel' data it generates can be resampled from the empirical data themselves, and the technique is robust. Both resampled and simulated parallel data do not yield substantively different results (Revelle, 2015). Moreover, PA is flexible, having been modified and improved upon since its conception, and is capable of assessing factor and component structures, as well as both ratio and ordinal data (Garrido, Abad, & Ponsoda, 2013). Finally, PA is noteworthy when contrasted with other, modern factor number tests because unlike even the best alternatives, e.g. comparison data (Ruscio & Roche, 2012), it is completely unbiased (cf. Courtney, 2013). Based on simulations, PA identifies the correct number of factors in more than 76% of cases (Ruscio & Roche, 2012). For this reason, it remains one of the best tests available for component/factor extraction.

All methods of course have their drawbacks (Ruscio & Roche, 2012); there is no 'one size fits' all approach. Even if some methods are demonstrably more accurate than others, e.g. PA versus Kaiser's criterion, few data sets will produce an immediate and clear solution. Therefore, it is paramount that no single automated extraction test be used as the sole method to determine how many components/factors to extract from a data reduction analysis. Instead, multiple automated tests should be implemented and compared. If multiple tests agree on the same number of components/factors to extract, then researchers can be confident with their decisions about extraction (Gorsuch, 1983).

<H1>What if multiple automated methods disagree?

It is not uncommon for multiple automated methods to disagree on the number of components to extract. As previously noted, in such cases a scree test may be used as a quick and easy ‘tie-breaker’ if the plot reveals a clear and distinct drop in the eigenvalues past a certain component/factor. Such instances, however, are becoming increasingly rare as automated methods are improved upon. Where appropriate, researchers should use PA as a tie-breaker because it is a robust technique, but we again caution readers to consider as many options as possible before settling on a particular selection of factors. For example, other sophisticated analyses such as Everett’s tests may be required to determine which model to use for subsequent analyses after extracting multiple solutions with differing numbers of factors (Everett, 1988).

Researchers should always keep in mind the theory they wish to test, and where theory is well established, it can be used to guide choices in how many factors to extract. If the analysis is wholly exploratory, or theories are at odds, there is nothing wrong with extracting multiple factor structures and comparing them when multiple extraction methods disagree on how many to extract. Factor interpretability can be assessed after extraction, and, depending on what variables are of interest, investigating additional associations may indicate which structure is the most useful (Altschul, Terrace, & Weiss, 2016). As with any model, however, researchers must beware of post hoc modification since greater degrees of freedom can hinder the generalizability of an analysis. Ideally, researchers should always keep their theory in mind throughout the analytical process, and factor solutions that are extracted should be interpretable in light of theory.

Finally, basic EFA or PCA may not be the best method for all situations. More complex and potentially hierarchical data may require a more advanced modelling approach. For example, EFA is itself a specific implementation of a more general SEM framework, which allows users to specify latent variables and all paths between latent and measured variables. If one suspects that a one-level factor model is not sufficient to explain the data, for example if there are unambiguous sources of nonindependence such as correlated error structure, then SEM should be considered because it is better suited for handling complex structures (Reise, Schneines, Widaman, & Haviland, 2013).

Ultimately, researchers need to be aware of what EFA and PCA are creating: reduced data that are only the result of what one has fed into one's analysis. Variable reduction may make data more manageable and possibly more interpretable, but the results are derived from noninferential matrices of correlations between variables, and there is no guarantee that these techniques will produce quantitatively superior data. The results of data reduction are contingent on the input; some data will be appropriate for data reduction, some simply will not. Moreover, similar but distinct data will yield different results. Comparing different data sets in the same or similar models is fundamentally qualitative, and researchers must bear this in mind when considering what to conclude from their analyses.

<H1>Performing and interpreting automated extraction tests in R

The following instructions are specific to the R programming language because of its wide use and robust, well-maintained feature set. All commands are available from base R, or the ‘psych’ package (Revelle, 2015). The code for running these analyses can be found in the Appendix.

First, data should be organized in a ‘data.frame’ format, which is native to R. We will call our example data.frame: ‘df’. The first column of the data.frame should contain the names of individuals and/or dyads. Many functions require only numerical input, and the first column can be subset out of the data.frame with the command ‘df[,-1]’. For example, to examine the correlation matrix of the data for suitability, the entire command ‘cor(df[,-1])’ will display the numeric correlation matrix. We also suggest using ‘corPlot’ in the same way, to view the correlation matrix graphically. Two specific tests for factorability, Barlett’s test and the Kaiser-Meyer-Olkin measure, can be found in psych and accessed using ‘cortest.bartlett(df[-1])’ and ‘KMO(df[-1])’.

Executing the command ‘nfactors(df[,-1])’ will display graphical representations of VSS, eBIC and SRMR (e.g. Fig. 2). It will also generate a myriad of other fitted statistics, which may be useful to the advanced user. Executing ‘fa.parallel(df[,-1])’ will display a plot, as in Fig. 3, as well as give a specific recommendation for how many components to retain for extraction.

As previously mentioned, EFA and PCA often produce very similar solutions in practice, but the underlying matrix algebra differs such that when each procedure is repeated, the results can differ considerably. Thus, while the other five extraction methods that we previously discussed need not distinguish between factors and components, PA must be adjusted to support EFA (Revelle, 2015).

In Fig. 2, the VSS test suggests that a three-factor model has a better fit than a one- or two-factor solution, meaning the three-factor model shows an improvement in fit over the one- and two-factor models, which is evident because the number three in the plot is above the line associated with the other two models. The empirical BIC test suggests two factors should be extracted since that model shows the lowest BIC compared to the others. The SRMR test indicates that models with two or more factors are acceptable.

In Fig. 3, based on Kaiser's criterion these artificial data cluster onto a single factor. By contrast, the scree plot suggests two factors, since the line appears to asymptote after the second eigenvalue. Similarly, the parallel analysis suggests extracting two factors, which is evident because the line representing the 'FA actual data' crosses the line representing the 'FA resampled data' after the two-point mark along the x-axis, i.e. those factors that are greater than the 95th percentile of the distribution of eigenvalues generated from the resampled data.

Collectively, based on this example, extracting two factors appears to be the most reasonable decision to make for a data reduction analysis since (1) half the automated tests, including parallel analysis (i.e. the most robust method), point towards a two-factor solution, (2) the SRMR test indicates that this decision is acceptable, and (3) the scree plot (i.e. our 'tie-breaker') corroborates this decision.

<H1>Summary and Future Directions

Data reduction analyses provide a unique and objective means through which researchers can interpret animal data, and the work that has already been done in this area has taken a very important step in that direction. With the increasing number of studies using this approach, researchers must take into careful consideration both the data reduction technique (PCA or FA) and the extraction method(s) used to reduce the number of components/factors within their data set. Failure to do this can have consequences in terms of comparability, replicability and interpretation of those results. In light of the well-known deficiencies associated with Kaiser's criterion, we emphasize that animal researchers must refrain from using this technique in future work and instead use more robust and automated extraction techniques (e.g. PA, empirical BIC, VSS, comparison data). If these automated tests recommend the same number of components/factors, then researchers can be confident about their decisions to extract. If they disagree, then as we discussed, there are multiple avenues to take to aid decision making on extraction and modelling frameworks. Avoiding Kaiser's criterion and supplementing scree tests with more robust and automated tests will greatly improve the utility and reliability of data reduction techniques, particularly for comparisons across studies. Of the methods we have discussed, we recommend PA and BIC in particular because of their strong performance under simulation (Ruscio & Roche, 2012), but novel methods are being developed with surprising frequency, and we encourage readers to explore the literature for newly verified methods.

Declaration of Interest

Both authors declare no conflict of interest.

Acknowledgments

We thank Dr. Alexander Weiss for fruitful discussion, and the referees for their useful feedback on the manuscript.

References

Adamo, S.A., Kovalko, I., & Mosher, B. (2013). The behavioural effects of predator-induced stress responses in the cricket (*Gryllus texensis*): the upside of the stress response. *Journal of Experimental Biology*, *216*, 4608–4614.

Altschul, D., Terrace, H., & Weiss, A. (2016). Serial Cognition and Personality in Macaques. *Animal Behavior and Cognition*, *3*, 46–64.

Arden, R., & Adams, M. J. (2016). A general intelligence factor in dogs. *Intelligence, 55*, 79-85.

Browne, M.W. (2001). An overview of analytic rotation in exploratory factor analysis. *Multivariate Behavioral Research*, *36*, 111–150.

Brown, J.D. (2009). Principal components analysis and exploratory factor analysis - Definitions, differences, and choices. *Statistics*, *13*, 26–30.

Cattell, R.B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276.

Costello, A.B., & Osborne, J.W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation*, *10*, 1–9.

Courtney, M.G.R. (2013). Determining the number of factors to retain in EFA: Using the SPSS R-Menu v2. 0 to make more judicious estimations. *Practical Assessment*, *Research & Evaluation*, *18*, 1–14.

Dunham, A.E., Maitner, B.S., Razafindratsima, O.H., Simmons, M.C., & Roy, C.L. (2013). Body size and sexual size dimorphism in primates: influence of climate and net primary productivity. *Journal of Evolutionary Biology*, *26*, 2312–2320.

Everett, J. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research*, *18*, 197–218.

Fabrigar, L.R., Wegener, D.T., MacCallum, R.C., & Strahan, E.J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods*, *4*, 272.

Ferreira, R. G., Mendl, M., Guilherme, P., Wagner, C., Araujo, T., Nunes, D., & Mafra, A. L. (2016). Coping strategies in captive capuchin monkeys (*Sapajus* spp.). *Applied Animal Behaviour Science*, *176*, 120–127.

Field, A. (2009). Discovering statistics using SPSS, 3rd edn. London, U.K.: Sage.

De Medeiros Filho, S.A., de Carvalho-Neto, F.G., Garcia, A.C.L., Montes, M.A., & Duarte-Neto, P.J. (2018). Morphometric variability in *Artibeus planirostris* (*Chiroptera: Phyllostomidae*) in environments with different states of conservation in the Atlantic Forest, Brazil. *Mammalian Biology*, *90*, 66–73.

Finger, N.M., Bastian, A., & Jacobs, D.S. (2017). To seek or speak? Dual function of an acoustic signal limits its versatility in communication. *Animal Behaviour*, *127*, 135–152.

Fraser, O.N., & Bugnyar, T. (2010). The quality of social relationships in ravens. *Animal Behaviour*, *79*, 927–933.

Fraser, O.N., Schino, G., & Aureli, F. (2008). Components of Relationship Quality in Chimpanzees. *Ethology*, *114*, 834–843.

Garrido, L.E., Abad, F.J., & Ponsoda, V. (2013). A new look at Horn’s parallel analysis with ordinal variables. *Psychological Methods*, *18*, 454–474.

Gorsuch, R.L. (1983). Factor analysis, 2nd edn. Hillsdale, NJ: L. Erlbaum.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103*, 265–275.

Hassrick, J.L., Crocker, D.E., & Costa, D.P. (2013). Effects of maternal age and mass on foraging behaviour and foraging success in the northern elephant seal. *Functional Ecology*, *27*, 1055–1063.

Hooper, D., Coughlan J., & Mullen, M.R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods*, *6*, 53–60.

Horn, J.L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*, 1–55.

Jolliffe, I.T. (1986). Principal component analysis and factor analysis. In Jolliffe (Ed.), *Principal component analysis* (pp. 115-128)., New York, NY: Springer

Kaiser, H.F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*, 141–151.

Keagy, J., Savard, J.-F., & Borgia, G. (2011). Complex relationship between multiple measures of cognitive ability and male mating success in satin bowerbirds, *Ptilonorhynchus violaceus*. *Animal Behaviour*, *81*, 1063–1070.

Khargharia, G., Kadirvel, G., Humar, S., Doley, S., Bharti, P.K., & Das, M. (2015). Principal component analysis of morphological traits of assam hill goat in eastern Himalayan India. *Journal of Animal and Plant Sciences*, *25*, 1251–1258.

Klein, S., Pasquaretta, C., Barron, A.B., Devaud, J.-M., & Lihoreau, M. (2017). Inter-individual variability in the foraging behaviour of traplining bumblebees. *Scientific Reports*, *7*, 4561.

Koski, S.E., De Vries, H., Van de Kraats, A., & Sterck, E.H. (2012). Stability and Change of Social Relationship Quality in Captive Chimpanzees (*Pan troglodytes*). *International Journal of Primatology*, *33*, 905–921.

Lawrence, M., Mastromonaco, G., Goodrowe, K., Santymire, R.M., Waddell, W., & Schulte-Hostedde, A.I. (2017). The effects of inbreeding on sperm morphometry of captive-bred endangered mammals. *Canadian Journal of Zoology*, *95*, 599–606.

Martin, J. G. A., & Reale, D. (2008). Temperament, risk assessment, and habituation to novelty in eastern chipmunks, *Tamias striatus*. *Animal Behaviour*, 75, 309-318.

McFarland, R., & Majolo, B. (2011). Exploring the Components, Asymmetry and Distribution of Relationship Quality in Wild Barbary Macaques (*Macaca sylvanus*). *PLoS One*, *6*, e28826.

Menzies, A. K., Timonin, M. E., McGuire, L. P., & Willis, C. K. R. (2013). Personality variation in little brown bats. *PLoS One*, *8*, e80230.

Meulman, E. J.M., & van Schaik, C.P. (2013). Orangutan tool use and the evolution of technology. In: Tool use in animals: cognition and ecology, eds. Crickette Sanz, Josep Call and Christophe Boesch. Cambridge, UK: Cambridge University Press, 176–202.

Moreno, K.R., Highfill, L., & Kuczaj, S.A. (2017). Does personality similarity in bottlenose dolphin pairs influence dyadic bond characteristics? *International Journal of Comparative Psychology*, *30*, 1–15.

Morton, F.B., Todd, A.F., Lee, P., & Masi, S. (2013). Observational monitoring of clinical signs during the last stage of habituation in a wild western gorilla group at Bai Hokou, Central African Republic. *Folia Primatologica*, *84*, 118–133.

Murray, A. L., & Johnson, W. (2013). The limitations of model fit in comparing the bi-factor versus higher order models of human cognitive ability structure. *Intelligence*, *41*, 407– 422.

Nath, A., Singha, H., Deb, P., Das, A.K., & Lahkar, B.P. (2015). Nesting in a crowd: response of house sparrow towards proximity to spatial cues in commercial zones of Guwahati City. *Proceedings of the Zoological Society*, 69, 249–254.

Poinapen, D., Konopka, J.K., Umoh, J.U., Norley, C.J.D., McNeil, J.N., & Holdsworth, D.W. (2017). Micro-CT imaging of live insects using carbon dioxide gas-induced hypoxia as anesthetic with minimal impact on certain subsequent life history traits. *BMC Zoology*, *2*, 1–13.

Posada, D., Buckley, T.R., & Thorne, J. (2004). Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests. *Systematic Biology*, *53*, 793–808.

Pritchard, A.J., Sheeran, L.K., Gabriel, K.I., Li, J.-H., & Wagner, R.S. (2014). Behaviors that predict personality components in adult free-ranging Tibetan macaques *Macaca thibetana*. *Current Zoology*, *60*, 362–372.

Rebecchini, L., Schaffner, C.M., & Aureli, F. (2011). Risk is a Component of Social Relationships in Spider Monkeys. *Ethology*, *117*, 691–699.

Reise, S.P., Scheines, R., Widaman, K.F., & Haviland, M.G. (2013). Multidimensionality and structural coefficient bias in structural equation modeling: A bifactor perspective. *Educational and Psychological Measurement*, *73*, 5–26.

Revelle, W. (2015). psych: Procedures for Personality and Psychological Research. Evanston, IL: Northwestern University. http://CRAN.R-project.org/package=psych Version = 1.5.4. Accessed 26 May 2015

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research*, *14*, 403–414.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological Assessment*, *24*, 282–292.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*, 461–464.

Slipogor, V., Gunhold-de Oliveira, T., Tadic, Z., Massen, J.J.M., & Bugnyar, T. (2016). Consistent inter-individual differences in common marmosets (*Callithrix jacchus*) in boldness-shyness, stress-activity, and exploration-avoidance. *American Journal of Primatology*, *78*, 961–973.

Stevens, J.M., De Groot, E., & Staes, N. (2015). Relationship quality in captive bonobo groups. *Behaviour*, *152*, 259–283.

Todorov, H., Fournier, D., & Gerber, S. (2018). Principal components analysis: theory and application to gene expression data analysis. *Genomics and Computational Biology*, *4*, e100041–e100041

Velicer, W.F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Velicer, W.F., & Jackson, D.N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research*, *25*, 1–28.

Venturini, G.C., Savegnago, R.P., Nunes, B.N., Ledur, M.C., Schmidt, G.S., El Faro, L., & Munari, D.P. (2013). Genetic parameters and principal component analysis for egg production from White Leghorn hens. *Poultry Science*, *92*, 2283–2289.

Willems, E.P., Arseneau, T.J.M., Schleuning, X., & van Schaik, C.P. (2015). Communal range defence in primates as a public goods dilemma. *Philosophical Transactions of the Royal Society: Biological Sciences*, *370*, 20150003.

Yakubu, A., & Okunsebor, S.A. (2011). Morphometric differentiation of two Nigerian fish species (*Oreochromis niloticus* and *Lates niloticus*) using principal components and discriminant analysis. *International Journal of Morphology*, *29*, 1429–1434.

Yong, A.G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*, 79–94.

Zwick, W.R., & Velicer, W.F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin*, *99*, 432–442.

Appendix

Here we give the code for performing automated extraction tests in R (Revelle 2015).

library(psych) ## Main package used in this annex.

```
require(GPArotation) ## Supplementary package - useful for rotations.

## Users should import their data set here, saving as 'df'.

### Inspecting the correlations between variables before testing.
cor(df[,-1]
    , use = 'pairwise.complete.obs' ## Default is 'everything' - can produce many NAs.
)

corPlot(df[,-1]) ## Graphical plot of the correlation matrix.

### Testing the suitability of the data for factoring.
cortest.bartlett(df[,-1]) ## Bartlett's test that the correlation matrix is the ID matrix.
## The P value should be low, indicating that correlations are not all 1, and multiple
## factors could be extracted.

KMO(df[,-1]) ## Kaier, Meyer, Olkin measure of sampling adequacy.
## Less than 0.5 for an item has been labelled unacceptable,
## but higher values (e.g. > 0.8) are generally preferred.

### Determining the number of factors to extract.
nfactors(df[,-1] ## Replicates the style of Fig. 2.
         , n = 10 ## Sets the maximum number of factors to search for - default is 20.
```

```
        , rotate = 'oblimin' ## Default is 'varimax' - an orthogonal rotation.
)
## Output plot shows VSS, eBIC, SRMR and Complexity (a general diagnostic statistic).
## Full output is displayed in the console, and additional statistics can be explored
## and plotted, e.g.:
plot(nfactors(df[,-1], n=10, rotate='oblimin')$map, type = 'b')
## Velicer's minimum average partial (MAP), which indicates the optimal number of factors
## where it reaches a minimum.

## To fully take advantage of the many nfactors statistics, we strongly recommend
## that users consult the help file:
?nfactors

## Parallel analysis of factors solutions.
fa.parallel(df[,-1]
        , sim = FALSE ## Default is TRUE - FALSE replicates style of Fig. 3.
        , SMC = FALSE  ## Ensures that PA is adjusted for factors.
        , fa = 'fa' ## Plots only the factor analyses.
)
## This plots a scree plot with adjusted eigenvalues and the data for comparison,
## which are random and/or resampled. Where the adjusted eigenvalue for a given factor
## is above the line of eigenvalues from random/resampled data, parallel analysis
## indicates that that factor ought to be retained.
```

Figure Captions

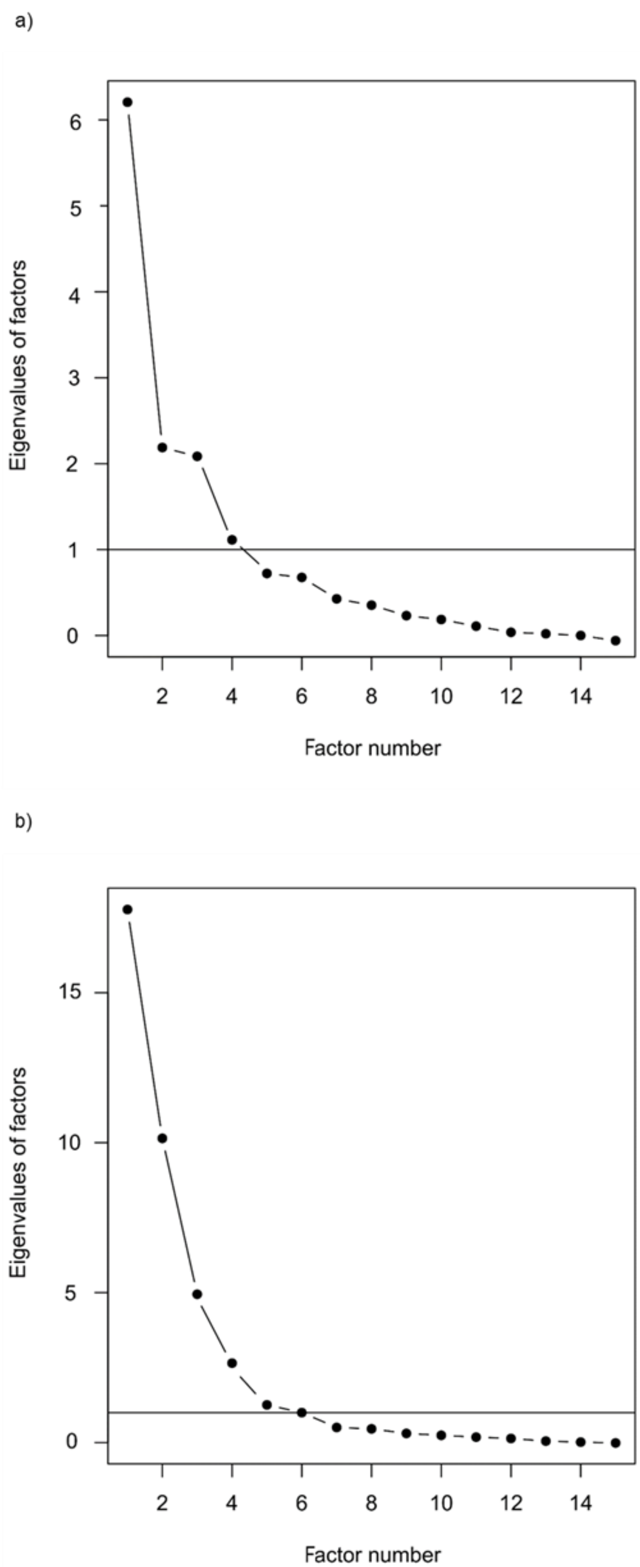

Figure 1. Example of scree tests on (a) clearly and (b) ambiguously factorable data sets.

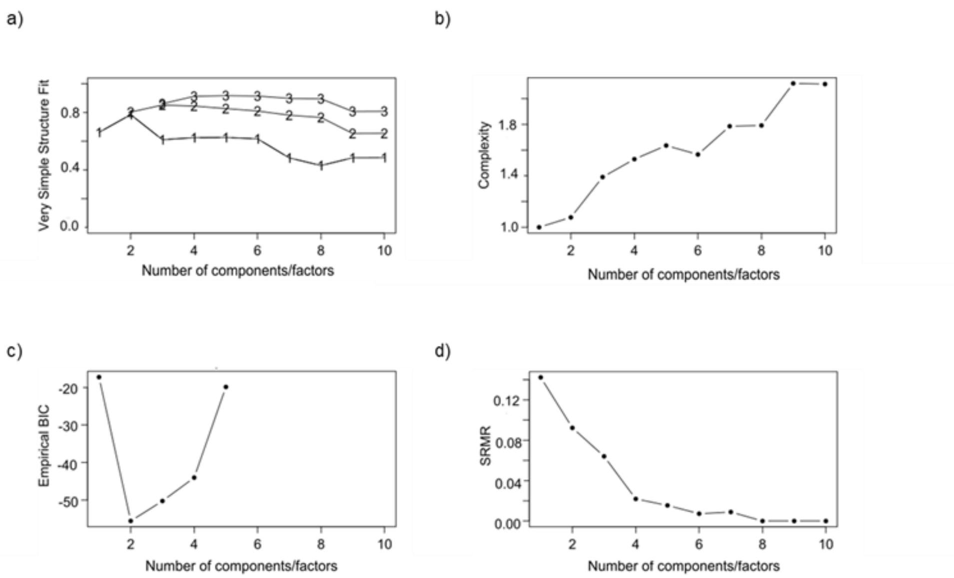

Figure 2. Example of plotted results using the R psych package ‘nfactors’ function, including (a) very simple structure, (b) complexity, (c) empirical BIC and (d) standardized root mean square residuals (SRMR). For the empirical BIC output, the number of variables (10) limits the calculation of empirical BIC to solutions of at most five components/factors.

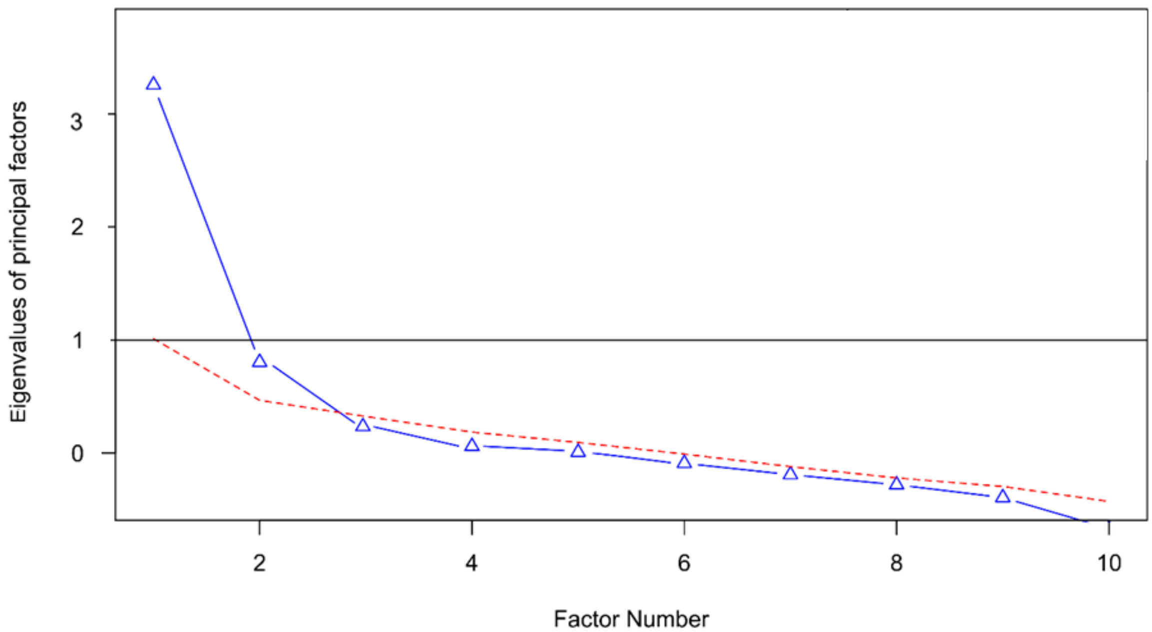

Figure 3. Example of results of parallel analysis, on a scree plot. Triangles represent eigenvalues generated from the actual data. Dashed lines represent random simulated eigenvalues. The horizontal black line at 1 represents Kaiser's criterion.