Multivariate methods and small sample size: combining with small effect size

Sergey V. Budaev

A.N. Severtsov Institute of Ecology and Evolution, Russian Academy of Sciences

Running headline: **Multivariate methods and small sample size**

Correspondence:
Dr. Sergey. V. Budaev,
A.N. Severtsov Institute of Ecology and Evolution,
Russian Academy of Sciences, Leninsky prospect 33,
Moscow 119071, Russia
Tel. (mobile): +7 985 456 32 24
Tel. (work): +7 495 952 40 17
E-mail: sbudaev@gmail.com

Date produced: 1 December 2010

Notice: This manuscript is my response to:

Dochtermann, N.A. & Jenkins, S.H. Multivariate methods and small sample sizes, Ethology.

and accompanies this paper:

Budaev, S. Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. **Ethology**, 116, 472-480 [abstract / full text].

In a recent paper, Dochtermann & Jenkins (2011) challenged the conventional views on sample size limitations in multivariate analysis. Using a series of simulations, they claim to demonstrate that a model comparison procedure can correctly rank alternative models in about 90% of cases with the sample size N=19. The authors generated random uncorrelated normal deviates, fitted a structural equation model assuming that all variables were independent and a similar model, assuming that all measures were intercorrelated due to an underlying latent construct. Comparison of the Akaike Information Criterion (AIC) fit indices of the two models revealed that the correct "null" model had to be chosen in the majority of cases over the full model. The authors concluded that the model-comparison approach (avoiding the null hypothesis significance testing) provides support for meaningful application of structural equation models in cases of very small effect size combined with small sample size (e.g. correlation coefficients about 0.2 with N=19, see Dochtermann & Jenkins 2007).

Theses results, however, are not as convincing as the authors suggest. First, about 50% of the model data were unusable because of non-convergence or other severe computational problems. This is a well known problem with small sample size (MacCallum et al. 1999; Boomsma & Hoogland 2001; Marsh, & Hau 1999). Second, the authors did not compare the null model with other possible models (e.g. a two-factor model), nor demonstrated that, with such small sample size, it is possible to correctly rank the "best" model if the manifest variables are not independent. They concede that the model comparison framework is different from the conventional null hypothesis significance testing and the Type I error rate is not applicable here. Unlike null hypothesis significance testing, the model selection paradigm involves weighting the evidence, rather than falsification of a hypothesis (see Anderson & Burnham 2002). It is also technically possible to compare the relative fit of two models neither of which adequately fits the data. Thus, the simulation does not provide much support for Dochtermann & Jenkins (2007).

Statistical inference is never conducted for its own sake, the researcher is interested to know to what degree the phenomenon can be replicated and predict other things. However, the combination of small sample size with low effect size does not allow either replication or prediction. To illustrate this, Figure 1 presents the values of the probability of replication p_{rep} for a range of Pearson correlation coefficients and sample sizes. p_{rep} is an estimate of the probability that a replication with the same power will give an effect of the same sign (see Killeen 2005). It is easy to see that small correlation coefficients are not replicable in cases of small sample size, making any models depending on these correlations not replicable and not generalizable too.

There exist various methodological obstacles to the application of multivariate methods in cases of small sample size, especially when it is combined with small effect size. Small samples may not provide a good representation of the general population. Accurate estimation of the model parameters in such cases is usually hardly possible. However, if the model parameters cannot be stabilized, comparisons between different models become very problematic: the researcher may simply find himself comparing wrong models. Third, samples can be strongly affected by outliers and measurement errors. But most robust estimation methods (e.g. Rousseeuw & Leroy 1987) are not applicable to small samples. Measurement error is likely to be a serious concern because observational studies are inherently prone to human mistakes (e.g. inaccurate timing, wrong identification and poor recognition of behavioural patterns etc., Vazire et al., 2007). Behavioural patterns are frequently characterized by significant flexibility, plasticity and stochasticity. Furthermore, adequate control of the environment in behavioural studies may be difficult, especially in the field. All this introduces various sources of error variability, which would affect small samples.

Fortunately, many of these problems can be rectified by increasing the effect size. There is a trade off between the effect size and sample size: stronger effect size allows for a smaller sample size and vice versa (see Taborsky 2010 for more discussion). Higher effect size also improves replicability and generalizability (e.g. see Fig. 1). Thus, it is not to say that small effect sizes are not scientifically important, they simply need larger sample size for valid inference. This is well known in the latent structure analysis. For example, Marsh & Hau (1999) recommend to have a minimum of 4 or 5 indicators with factor loadings (an estimate of effect size) exceeding at least 0.6 in confirmatory factor analysis when the simple size is small (see Budaev 2010 for more discussion). Animal behavioural researchers are aware of the more stringent sample size requirements in multivariate analysis: whereas the average N in animal behaviour research ranges between 20 and 30 (Taborsky 2010), factor analysis and PCA are on average based on *N*=64 (Budaev 2010).

Multivariate diagnostics

Dochtermann & Jenkins (2011) also challenged the usefulness of two classical diagnostic tests that are used prior to multivariate analysis: the Bartlett's sphericity test and the Kaiser-Meyer-Olkin measure of sampling adequacy (KMO). They cited several studies showing that the Bartlett's test has low statistical power when the average correlation in the correlation matrix is about 0.2 and the sample size is less than 90. They conclude that the utility of these indices is limited. This is a wrong argument: the only way to overcome the problem of insufficient power is to use appropriate sample size for the given effect size or improve the experimental design to increase the effect size (see Taborsky 2010 for more discussion).

Dochtermann & Jenkins (2011) also conducted simulations to determine the behaviour of the KMO in cases of small sample size and little effect size. Their conclusion is that KMO is not informative in behavioural research due to insufficient power to detect little correlations in cases of small N. The results of their simulations, however, seem to contradict their conclusion. With N=20 and average correlation 0.2, the average KMO exceeds the threshold value of 0.5 (see Fig 3 in Dochtermann & Jenkins 2011) defining the “unacceptable” calibration value (Dziuban and Shirkey 1974).

Small-sample behaviour of Bartlett's test and KMO

Prior studies investigated the power of the Bartlett's sphericity test and KMO in a relatively narrow ranges of N (e.g. N=20 and 200, Knapp & Swoyer 1967; Wilson & Martin 1983; N=50, 100, 250, 500, 1000, Dziuban et al. 1979), not well representing the range typical in animal behaviour research. Thus, I decided to perform a series of simulations aimed to assess the behaviour of these indices with small sample size. Using the R library mvtnorm (R Development Core Team 2008), I produced multivariate data sets in which random latent factors were linked with manifest variables. Thus, multivariate structures with explicit measurement and structural models were simulated instead of just matrices with specific average correlations (as Dochtermann & Jenkins 2011 did). To get a simplified overall picture, I generated structures with one latent factor and four manifest indicators (Model 1), one latent factor and eight manifest indicators (Model 2), and two latent factors linked with four manifest indicators (total eight indicators, Model 3). Factor loadings ranged from 0.1 to

0.8. In the Model 3, the two factors were uncorrelated. I also simulated random uncorrelated normal deviates with the same N (random data). All salient factor loadings L within a particular model were identical (e.g. in Model 1 it could be 0.3,0.3,0.3,0.3; Model 2, 0.3,0.3,0.3,0.3,0.3,0.3,0.3,0.3; in Model 3 with two factors: F1: 0.3,0.3,0.3,0.3,0.0,0.0,0.0,0.0, F2: 0.0,0.0,0.0,0.0,0.3,0.3,0.3,0.3). Sample size ranged from 10 to 100, reflecting the range typical in animal behaviour research.

The results of these simulations (Figure 1, left panel) agree that the Bartlett's sphericity test has low statistical power when the variables are weakly correlated. Clearly, with N=20, Bartlett's test fails to reject the null hypothesis (random matrix) in the majority of cases unless the manifest variables have high factor loadings (L>0.6, see Fig, 1). However, this level of factor loading is the minimum threshold for confirmatory factor analysis of small sample size (Marsh & Hau 1999) and minimum meaningful interpretation of the factors in exploratory factor analysis and PCA (see Budaev 2010 for more discussion). In most these analysis, measures with low loadings (L<0.5, or with large sample size L<0.4) are left uninterpreted. Thus, the Bartlett's sphericity test must be considered an adequate diagnostic tool within the normal range of factor analysis applications, especially given the magnitude of the average sample size used in animal behaviour research involving factor analysis N=64.

A notable feature of Fig. 1 is slight elevation of the probability of the null hypothesis rejection in data sets with low loadings (L<0.5) and N=10 observed when the number of manifest variables is high (Models 2 and 3, 8 variables). This reflects an inflation of the Type I error rate. It can be predicted that such detrimental effect can be exacerbated when small sample size is combined with large number of poorly correlated manifest variables.

The Kaiser-Meyer-Olin measure of sampling adequacy showed even better properties in these simulations (Fig, 1, right panel). KMO was very good at sorting out data sets with very low loadings (L<0.3 and random) in a wide range of sample sizes (N>20). Note that with the typical sample size N=60 and loadings L=0.6, KMO was within or exceeded the “middling” calibration range (KMO>0.7). Increasing the number of indicators per factor (see Fig. 1d) significantly increased KMO, which agrees with the classical recommendation (see Budaev 2010 for more discussion) to have several indicators for each factor. In contrast to the Bartlett's test, there was no indication of Type I error inflation.

Thus, contrary to Dochtermann & Jenkins (2011), both the Bartlett's test and the Kaiser-Meyer-Olkin measure of sampling adequacy proved to be quite informative within the range of conventional application of multivariate analysis. One important conclusion is that N=20 seems to represent the minimum sample size for the application of these multivariate methods because the power of both the Bartlett's test and the KMO are very quickly deteriorating below this threshold while the probability of Type I error may increase. With very small sample sizes, it becomes almost impossible to test the basic assumption of multivariate analysis, namely that the measures share common variance. Finally, my opinion on the data analysis presented by Dochtermann & Jenkins (2007) did not change. The sample size used in this study was too small for the effect size[1].

Literature Cited

Anderson, D.R. & Burnham, K.R. 2002: Avoiding pitfalls when using information-theoretic methods. J. Wildl. Management **66**, 912-918.

Boomsma, A. & Hoogland, J.J. 2001: The robustness of LISREL modeling revisited. In:

[1] Dochtermann & Jenkins (2011) also wonder what sample size was used in the random hypothetical example presented in Budaev (2010). The simulation used N=10. However the exact values of the PCA are unimportant in this example. What is important is that PCA of a random data matrix will produce spuriously interpretable factor pattern.

Structural Equation Modeling: Present and Future. (Cudeck, R., Du Toit, S. & Sorbom, D., eds). SSI Scientific Software, Chicago, pp. 139-168.

Budaev, S. 2010: Using principal components and factor analysis in animal behaviour research: Caveats and guidelines. Ethology **116**, 472-480.

Dochtermann, N.A. & Jenkins S.H. 2007: Behavioural syndromes in Merriam's kangaroo rats (*Dipodomys merriami*): A test of competing hypotheses. Proc. R. Soc. London B **274**, 2343-2349.

Dochtermann, N.A. & Jenkins S.H. 2011: Multivariate methods and small sample size. Ethology.

Dziuban, C.D., & Shirkey, E.S. 1974: When is a correlation matrix appropriate for factor analysis? Some decision rules. Psychol. Bull. 81, 358-361.

Dziuban, C.D., Shirkey, E.S. & Peeples, T.O. 1979: An investigation of some distributional characteristics of the measure of sampling adequacy. Educat. Psychol. Measur. **39**; 543-549.

Killeen, P. R. 2005: Replicability, confidence, and priors, Psychol. Sci **16**, 1009–1012.

Knapp, T.R. & Swoyer, V.H. 1967: Some empirical results concerning the power of Bartlett's test of the significance of a correlation matrix . Am. Educ. Res. J. **4**; 13-17.

MacCallum, R.C., Widaman, K.F., Zhang, S. & Hong, S. 1999: Sample size in factor analysis. Psychol. Meth. **4**, 84-99.

Marsh, H. W., & Hau, K. T. 1999: Confirmatory factor analysis: Strategies for small sample sizes. In: Statistical issues for small sample research. (Hoyle, R.H., ed.). Sage, Thousand Oaks, CA, pp. 251-284.

R Development Core Team 2008: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org.

Rousseeuw, P. J. & Leroy, A. M. 1987: Robust Regression and Outlier Detection; Wiley, Hoboken, NJ.

Taborsky, M. 2010: Sample size in the study of behaviour. Ethology **116,** 185–202 .

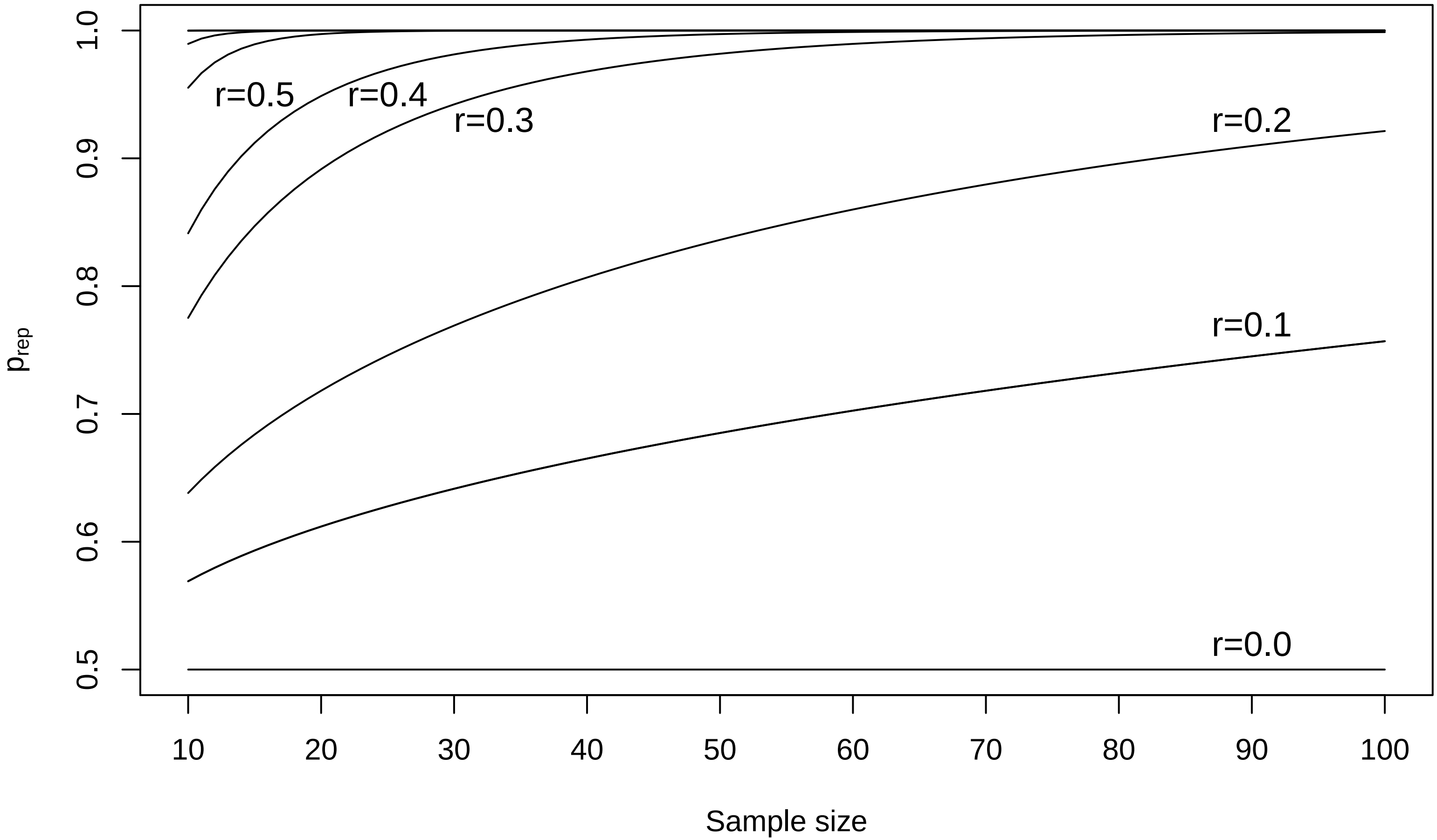
Vazire, S., Gosling, S.D., Dickey, A.S., & Schaprio, S.J. 2007: Measuring personality in nonhuman animals. In: Handbook of Research Methods in Personality Psychology (Robins, R.W., Fraley, R.C., & Krueger, R.F., eds.). Guilford, New York, pp. 190-206.

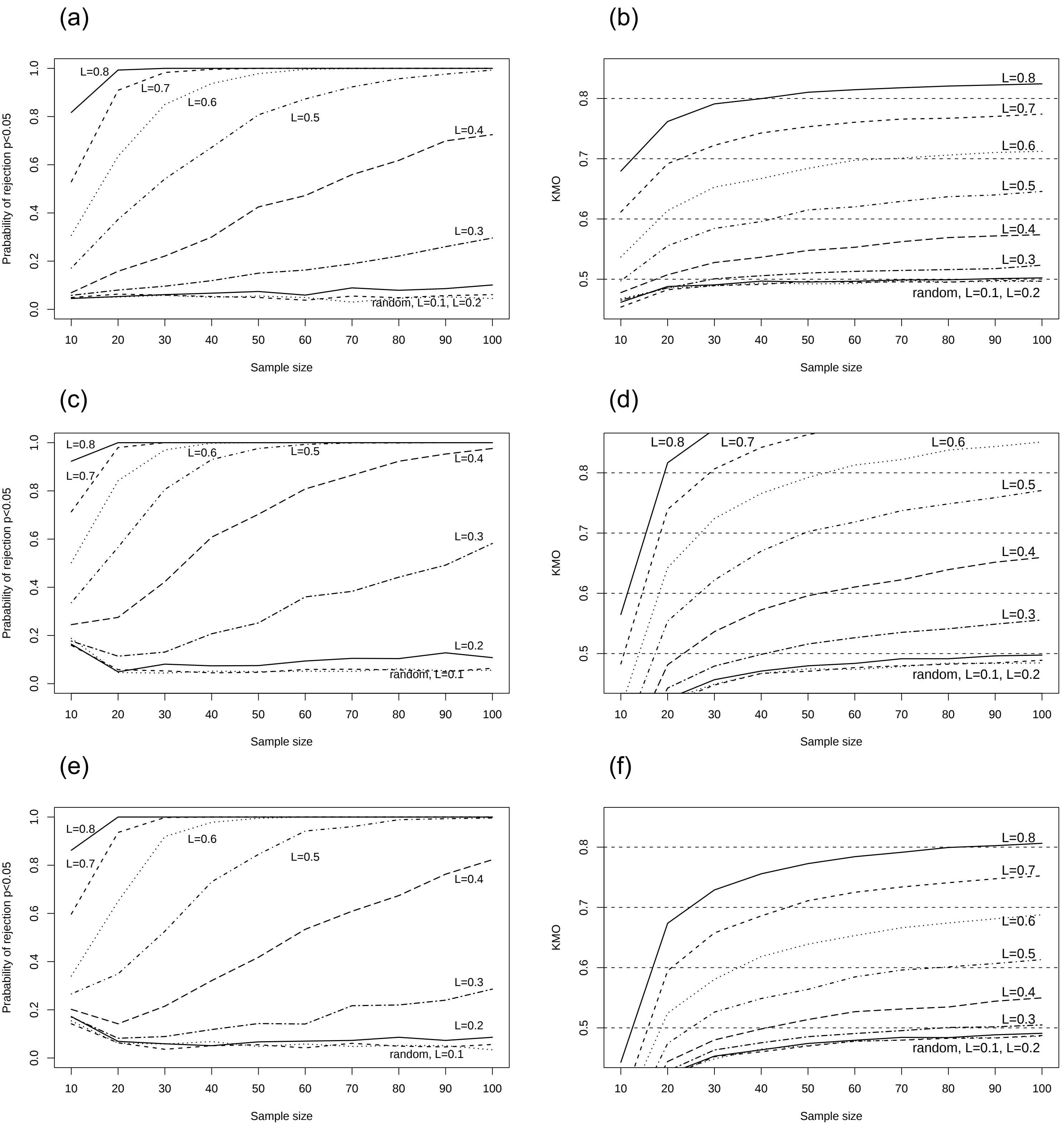
Wilson, G.A. & Martin , S.A. 1983: An empirical comparison of two methods for testing the significance of a correlation matrix , Educat. Psychol. Measur. **43,** 11-14.

Figure legend

Figure 1. Probability of replication p_{rep} of the Pearson correlation coefficients (ranging from r=0 to r=0.9) for a range of sample sizes (N=10-100). p_{rep} is an estimate of the probability that a replication with the same power will give an effect of the same sign and is thought to be an alternative of the conventional p-value (see Killeen, 2005).

Figure 2. Percentage of the cases when the Bartlett's sphreticity test rejected the null hypothesis wit $p<0.05$ (left panel) and the average value of the Kaiser-Meyer-Olkin measure of sampling adequacy (right panel) in various simulations. (a) and (b): Model 1, one latent factor, four manifest variables; (c) and (d): Model 2, one latent factor, eight manifest variables; (e) and (f): Model 3, two latent factors with eight manifest variables.


1.0
0.9
0.8
0.7
0.6
0.5
p_{rep}
r=0.5
r=0.4
r=0.3
r=0.2
r=0.1
r=0.0
10
20
30
40
50
60
70
80
90
100
Sample size


(a)
Prabability of rejection p<0.05
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
random, L=0.1, L=0.2
(b)
KMO
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
random, L=0.1, L=0.2
(c)
Prabability of rejection p<0.05
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
L=0.2
random, L=0.1
(d)
KMO
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
random, L=0.1, L=0.2
(e)
Prabability of rejection p<0.05
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
L=0.2
random, L=0.1
(f)
KMO
Sample size
L=0.8
L=0.7
L=0.6
L=0.5
L=0.4
L=0.3
random, L=0.1, L=0.2
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
10
20
30
40
50
60
70
80
90
100