

# Literature Search on the subject “Sample size in factor analysis”

Assembled by Jia Guo biostatistics grad student September 25, 2004

## ***1 Why sample size matters in factor analysis?***

It is widely understood that the use of larger sample sizes in applications of factor analysis tends to provide results such that sample factor loadings are more precise estimates of population loadings and are also more stable, or less variable, across repeated sampling.

## ***2 Published Guidelines***

Several approaches have been taken to propose guidelines for the sample sizes.

### ***(1) N:p ratio***

(N is the number of subjects, p is the number of observed variables)

- Idea :  
The rules relating N to p seem to be based on the shrinkage concept developed in multiple regression.
- Reference  
Baggaley, 1982 ; Brislin, Lonner, & Thorndike, 1974;  
Cattell, 1952, 1978; Gorsuch, 1983 (5:1);  
Hair, Anderson, Tatham, & Grablovsky, 1979;  
Kunce, Cook, & Miller, 1975;  
Lindeman, Merenda, & Gold, 1980;  
Marascuilo & Leven, 1983; Nunnally, 1978 (10:1)
- Guideline  
Suggested *N-to-p* ratios are varying from 2:1 to 20:1.

### ***(2) Absolute N***

- Idea  
The recommendation for a minimum sample size of 100 to 200 observations is probably based on the argument that a correlation coefficient becomes an adequate estimator of the population correlation coefficient when sample sizes reach this level.
- Reference  
Comrey. 1973, 1978; Gorsuch, 1983; Guilford, 1954;  
Hair et al., 1979; Lindeman et al., 1980; Loo, 1983).

- **Guideline**  
A minimum N of 100 to 200 observations is also often recommended.
- Using real data to empirically test the relation between sample size and the stability of the sample solution.

Aleamoni (1973); Barrett and Kline (1981)  
 Arrindell and van der Ende (1985);  
 Velicer, Peacock, and Jackson (1982).

**(3) *N:m ratio*** (m is the number of expected factors)

(Cattell, 1978).

All previous empirical investigations provide results that are limited in scope, focusing on one data set or one value of p;  
 The most familiar advice given the researcher, however, is to obtain the maximum sample size possible (Guertin & Bailey, 1970; Humphreys, IIgen, McGrath, & Montanelli, 1969; Press, 1972; Rummell, 1970).

**(4) *Necessary N depends on several specific aspects of a given study***

Browne(1969) P:m ratio; Pennell(1968) communality; Velicer and Fava(1998) communality;

### **3 Simulation study**

- (1) Guadagnoli and Velicer (1988) concluded that the component saturation and absolute sample size were the most important factor in determining stability. To a lesser degree, the number of variables per component was also important, with more variables per component producing more stable results. (PCA)
- (2) Jackson (2001)'s Monte Carlo investigation didn't find the significant effect of the ratio.

#### 4 Theoretical work and simulation study

Two papers by MacCallum et.al (1999&2001)

- Idea

These two papers pointed out:

“A fundamental misconception about this issue is that the minimum sample size or the minimum ratio of sample size to number of variables is invariant across studies. In fact necessary sample size is dependent on several aspects of any given study, including the level of communality of the variables and the level of overdetermination of the factors.”

- Conclusions

The first paper considered the cases only with sampling error and no model error. The authors concluded that “ Sample size and level of overdetermination had little effect on the recovery of population factors when communalities were high; Only when some or all of the communalities became low did sample size and overdetermination become important determinants of recovery of population factors.”

There is also significant interactive effects of overdetermination with communality level and sample size.

The second paper considered the cases with both two types of errors and concluded that “ the lack of fit of the model in the population will not, on average, influence recovery of the population factors in the analysis of sample data, regardless of degree of model error and regardless of sample size. Rather, such recovery will be affected only by phenomena related to sampling error which have been studied previously”.

- Guideline

With communalities in the range of 0.5, it is still not difficult to achieve good recovery of population factors, but one must have well determined factors and possibly a somewhat larger sample, in the range of 100-200

When communalities are consistent low, but there is high overdetermination of factors (6, 7 indicators per factor and a rather small number of factors), one still can achieve good recovery of population factors, but larger samples are required- probably over 100.

With low communalities, a small number of factors, and just three or four indicators per factor, a much larger sample is needed- probably at least 300.

The worst case-low communalities and a larger number of weakly determined factors, any possible good recovery requires very large samples, well over 500.

- Limitations

1. Simulation only studies limited range of  $N$ ,  $p$ ,  $r$
2. This approach only focuses on the particular objective of obtaining solutions that are adequately stable and congruent with population factors. May not satisfy other objectives.
3. Sampling error should influence solutions in CFA.

**Sample size in factor analysis.**

By MacCallum, Robert C.; Widaman, Keith F.; Zhang, Shaobo; Hong, Sehee (1999) *Psychological Methods*. 4(1), 84-99.

**Sample size in factor analysis: The role of Model Error**

By MacCallum, Robert C.; Widaman, Keith F.; Preacher, Kristopher J.; Hong, Sehee (2001) *Multivariate Behavioral Research*. 36(4), 611-637.