The Technical Whitepaper Series

4

Factor Stability

numbers of variables, cases, and factor-solution stability

pbarrett.net
Advanced Projects R&D

April, 2005

The relationship between sample size, scale, and factor stability

Sample size is a critical factor which affects estimation accuracy for any statistical entity, including population distribution parameters such as the mean, median, and standard deviation and other statistical parameters such as correlation coefficients, beta weights in regression, test scale reliability, and item and scale factor loading patterns. Sample size is also a critical factor within statistical power analysis; that is, estimation of the sample size required to correctly reject the null hypothesis at a particular threshold probability (alpha) within a statistical significance-test process. The very meaning of the word "sample" suggests that the observations made within it are to be considered in as representative of those to be found in a wider "population" of all possible observations of a particular variable. By taking a subset (sample) of such observations, we immediately encounter potential problems of **bias** and **accuracy** of parameter estimation. Bias is apparent where we collect observations in such a way that our sample consists of particular ranges and frequencies of observations which are not representative of the total range and frequencies of observations expected to be constitutive of our hypothesised population. Accuracy is compromised where we collect too few observations of a variable's possible values such that the estimation of the parameters of interest is flawed. Obviously, sample bias is reflected in parameter estimation accuracy. However, we may collect a relatively unbiased sample which nevertheless consists of too few observations to allow computation of an accurate estimate of a parameter.

Sample size always plays a part in the calculation of variability of a parameter estimate. **In general, the fewer observations made of a variable, the larger the variability of an estimate of any parameters estimated from them**. This simple principle affects calculations of the variance and standard deviation, the standard error of a mean, correlation coefficients, reliability estimates, standard errors of measurement of test scores, and basically all statistical estimates which rely upon a measure of variability of a parameter as part of their calculation (as do all statistical significance tests). For example, if we take a small and large sample of observations in order to compute the mean of them, given all observations consist of the integers 1, 2, and 3 in equal proportion, and then estimate the sampling error of this mean, we see

Sample N	No. of 1s, 2s, and 3s	Mean	Standard Error of the Mean
9	1= 3	2.0	**0.287**
	2=3		
	3=3		
600	1=200	2.0	**0.033**
	2=200		
	3=200		

Likewise, if we compute the standard error of a correlation coefficient, to be used as part of the significance test of a null hypothesis that an observed correlation is significantly different from zero, we see that the sample size plays a critical role in the formula for computing a t-statistic:

$$t = \frac{r - \rho}{s_r}$$

where $s_r =$ the standard error of the correlation coefficient and

$t =$ the standardized value referenced against the t-distribution

$\rho =$ the hypothesised population correlation coefficient (= 0.0)

and

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}} \quad \text{where } n = \text{the sample size}$$

So, if we observe a correlation of 0.4, with samples sizes of 20 and 100, we would obtain:

For n=20

$$t = \frac{0.4 - 0.0}{s_r}$$

$$s_r = \sqrt{\frac{1 - 0.4^2}{20 - 2}} = \sqrt{\frac{0.84}{18}} = 0.216$$

$$t = \frac{0.4}{0.216} = \mathbf{1.8516}$$

For n=100

$$t = \frac{0.4 - 0.0}{s_r}$$

$$s_r = \sqrt{\frac{1 - 0.4^2}{100 - 2}} = \sqrt{\frac{0.84}{98}} = 0.0926$$

$$t = \frac{0.4}{0.0926} = \mathbf{4.32}$$

The correlation for n=20 would be adjudged “not significant” given a specified null hypothesis alpha of 0.05 two tail, as t=1.8516 has an associated two-tail probability of 0.081.

The correlation for n=100 would be adjudged “highly significant” given a specified null hypothesis alpha of 0.05 two tail, as t=4.32 has an associated two-tail probability of 0.00004.

When we come to the calculation of correlation coefficients and regression parameters, we have to be especially careful of the ratio of the numbers of cases of observations to the number of variables. For example, if we calculate a correlation between the scores on two tests, using just two individuals as our sample, the correlation will always be 1.0. Likewise a regression equation will always yield perfect prediction. Figures 1 and 2 below show why… doesn’t matter what value the variables take!

Figure 1

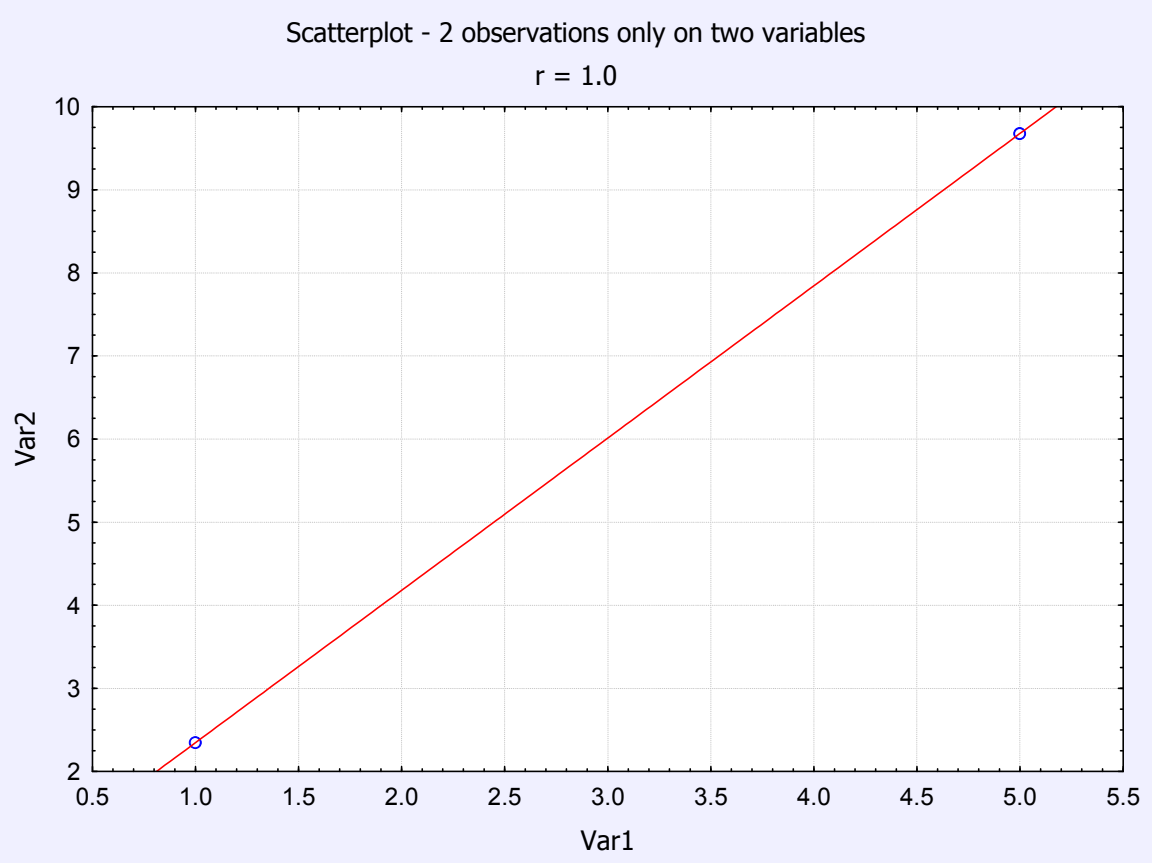
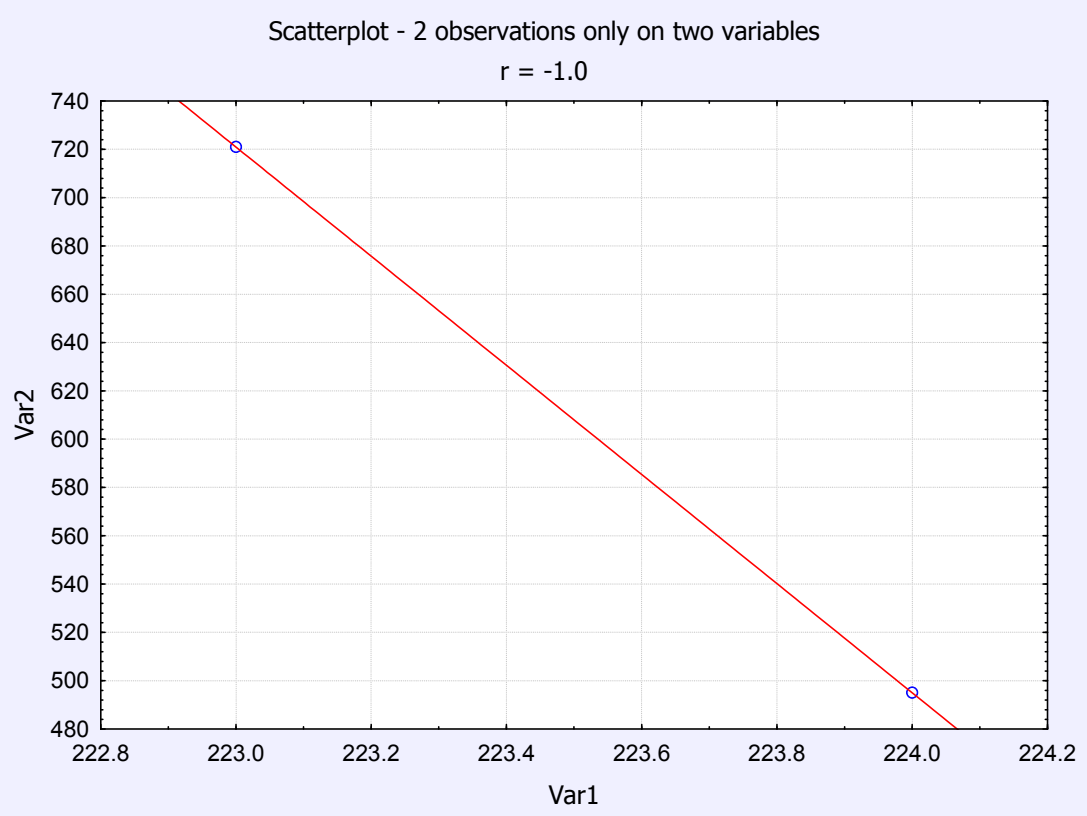
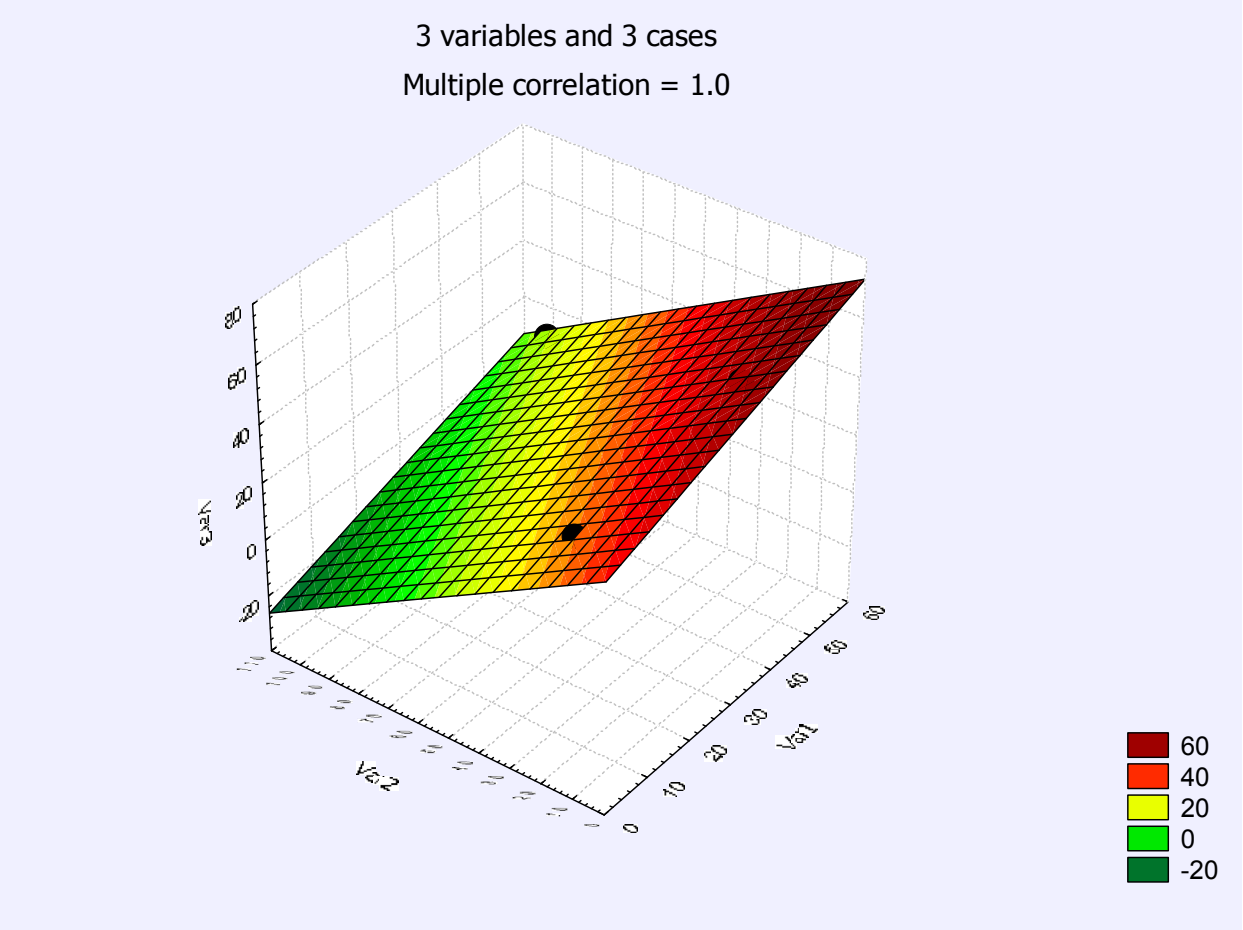

Scatterplot - 2 observations only on two variables
r = 1.0
Var2
Var1
10
9
8
7
6
5
4
3
2
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5

Figure 2


Scatterplot - 2 observations only on two variables
r = -1.0
740
720
700
680
660
640
620
600
580
560
540
520
500
480
Var2
222.8
223.0
223.2
223.4
223.6
223.8
224.0
224.2
Var1

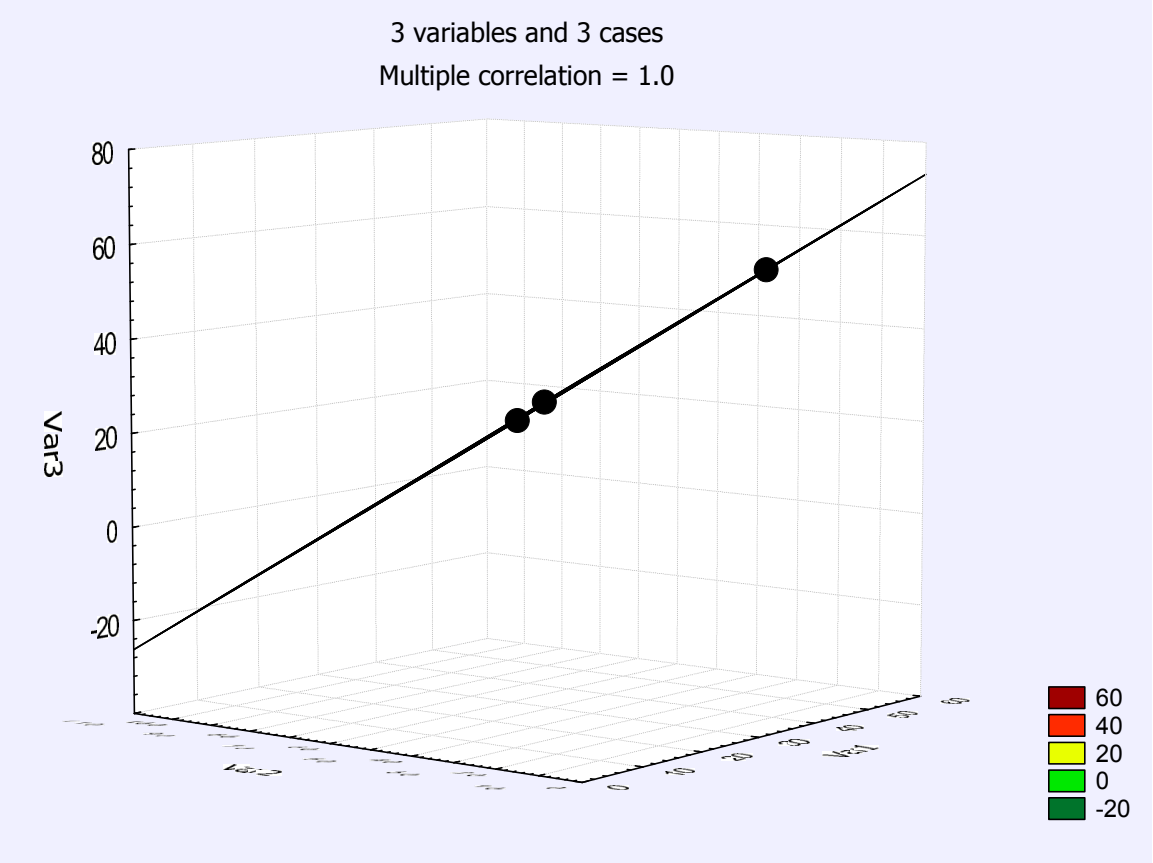
Now consider the case where we look at the relationship between three variables simultaneously (or where two values are required to predict a third), and we use data from only 3 individuals. Again, prediction will be perfect and the multiple correlation will be 1.0. Figures 3 and 4 below shows why …

Figure 3



As you can see from figure 3, because we are dealing with a 3-dimensional relationship, the “line” between points is actually a 3D-plane (or flat surface) rather than a 2D line. The black dots represent the three cases with test scores on the three variables. We can see the “line” if we look through all three dimensions along the angle of the plane

Figure 4



So, even if we had 80 cases, but using 80 variables, we are guaranteed to observe perfect prediction and a multiple correlation of 1.0.

What this demonstrates is that there will be a consistent bias upwards of any correlation coefficient (and regression prediction accuracy) as the ratio of cases to variables decreases (approaches 1:1). This is why the shrinkage formula is used for all multiple correlation coefficients.

For example, the formula for correcting a standard Pearson bivariate (2-variable) correlation is:

$$r_{adj} = \sqrt{1 - \frac{(1 - r^2)(n - 1)}{(n - 2)}} \qquad \text{where}$$

n = the sample size

r = the observed correlation

r_{adj} = the adjusted correlation (an unbiased estimate)

So, in our case where we had just 2 variables and 2 cases, and an observed correlation of 1.0, the adjusted correlation = 0.0 (actually non-computable). In the case where we observed a correlation of 0.4 with 20 cases, the adjusted correlation would be 0.34. for 100 cases, with the same 0.4 observed correlation, the adjusted value would be 0.39.

The adjustment made to a multiple correlation so as to estimate an unbiased value (taking into account the numbers of variables and sample size) is:

$$R_{adj} = \sqrt{R^2 - (1 - R^2)\left(\frac{m}{n - m - 1}\right)} \quad \text{where}$$

$R =$ the observed multiple correlation coefficient

$n =$ the sample size

$m =$ the number of "predictor" variables

$R_{adj} =$ the adjusted multiple correlation

This formula is the one used in SPSS, SAS, BMDP, and STATISTICA. The formula provided in Nunnally (1978, Eq. 5-23, p. 180)) differs only marginally from this one …

$$R^2_{adj} = 1 - (1 - R^2)\left(\frac{n - 1}{n - m}\right)$$

$$R_{adj} = \sqrt{R^2_{adj}}$$

The correction applied to the case where we had three cases and three variables (two predictor and one "dependent" or "outcome" variable), with an observed correlation of 1.0 yields …

$$R_{adj} = \sqrt{1.0^2 - (1 - 1.0^2)\left(\frac{2}{3 - 2 - 1}\right)}$$

where again we have a division by zero and so the result is fixed as 0.0.

But, assume we observe a multiple correlation of 0.65, using 10 variables as predictors, and a sample size of 35 cases. The corrected value is 0.43. If we observed a multiple correlation of **0.88** with 22 variables and 30 cases, **the corrected value would be 0.26**. Again, the importance of the sample size is readily appreciated here. If we had just doubled our sample size to 60, with the same correlation and variables, the adjusted R would now be 0.80, very near our observed value.

The correction or adjustment formula takes into account the ratio of cases to variables and attenuates the observed coefficient accordingly so as to compute an unbiased estimate. Indeed, the logic of this

correction is used by Nunnally (1978, p.180) to derive his well-known suggested ratio of 10 cases to every variable…

"if there are only 2 or 3 independent variables and no pre-selection is made among them, 100 or more subjects will provide a multiple correlation with little bias. In that case, if the number of independent variables is as large as 9 or 10, it will be necessary to have from 300 to 400 subjects to prevent substantial bias".

Green (1991), in a review paper entitled "How many subjects does it take to do a regression analysis" noted that ratios of cases to variables ranged from

- **5:1** (Tabachnik and Fidell (1989), p. 128-129)
- **10:1** to **40:1** (Tabachnik and Fidell, 2001, p. 180)
- **10:1** (Nunnally, 1978, p.180)

However, using the concepts of power analysis and effect size, Green showed that these simple rules of thumb were not very useful. Instead he proposed a simple formula which works fairly well where the number of predictor variables is less than 7, and a medium sized-multiple correlation of about 0.3 is expected:

$$N \geq 50 + 8m \qquad \text{where}$$

$m =$ the number of predictor variables

$N =$ the "to be used" sample size

So, with 5 predictor variables, the sample size should be a minimum of 90, which yields a ratio of **18:1**. With 2 predictor variables, the sample size should be a minimum of **33:1**. With m equal to or greater than 7 variables, Green proposed a more accurate and complex formula based upon *a priori* knowledge or hypothesis of the expected effect size (multiple correlation) …

$$N \geq \left(\frac{8}{f^2}\right) + (m-1) \qquad \text{where} \quad f^2 = \frac{R^2}{\left(1-R^2\right)}$$

$R^2 =$ the squared multiple correlation coefficient

$m =$ the number of predictor variables

So, with 10 predictor variables, and an expected multiple correlation of 0.5, the sample size should be 33. A case to variable ratio of about **3:1**. However, if we proposed an expected correlation of just 0.3, the sample size is computed as 90, which provides a case to variable ratio of **9:1**.

However, when we come to factor analysis and the stability of factor loadings, the rules of thumb concerning ratios of numbers of cases to variables appear to be of limited value. Indeed recent empirical evidence from studies which examine recovery of known factor structures at various sample sizes indicates that the stability of factor loadings is less a function of the sample size or ratio of cases to variables, and much more to do with the amount of "in-common" variance amongst variables available

to be explained by a set of factors. For example, Guadagnoli and Velicer (1988) developed a formula for estimating the average distance between a hypothesised population factor loading and its sample value, given a sample size. Velicer and Fava (1998) showed that this formula was applicable both to component and common factor analysis. The formula is:

$$Y = 1.10\left(\frac{1}{\sqrt{n}}\right) - 0.12(v) + 0.066 \qquad \text{where}$$

$Y =$ the average distance between a sample and population loading

$n =$ the sample size

$v =$ the hypothesised population loading

So, for a sample size of 100, with an average target loading of 0.6 on a factor, the likely discrepancy would be 0.104. For a sample size of 70, with an average target loading of 0.5 on a factor, the discrepancy would be 0.14. We can re-express the algebra here to allow us to compute the required sample size to attain a factor solution where we wish to observe a desired discrepancy between sample and hypothesised population factor loading:

$$n = \frac{302500}{\left(500(Y) + 60(v) - 33\right)^2}$$

for example, if we wish to observe a discrepancy at least as small as 0.05 between our hypothesised loading and our sample-based one, we would need a sample size of at least 625.

Further simulation work by MacCallum, Widaman, Zhang, and Hong (1999) and MacCallum, Widaman, Preacher, and Hong (2001) has indicated that this formula of Guadagnoli and Velicer is a fairly accurate representation of the relationship between sample size and factor loading, developing the understanding of "why" much further with an explanation couched in terms of variable "communalities" (the explained variation of a variable by one or more factors in a solution). In MacCallum et al (2001), the authors conclude …

"Within the context of exploratory factor analysis, our results have more relevance than do previous results for empirical studies because of the explicit incorporation of model error, which will always be present to some degree in practice. Our results also reinforce the evidence regarding the critical role of communality level, while demonstrating the limited value of traditional rules of thumb regarding sample size in factor analysis. For example, one common rule of thumb regarding sample size is that, to ensure stability of a factor solution, a researcher should have a ratio of subjects to variables that is 4:1 or larger. However, our results show that if communalities are high, recovery of population factors in sample data is normally very good, regardless of sample size, level of over-determination, or the presence of model error. Thus, samples somewhat smaller than traditionally recommended are likely sufficient when communalities are high. When communalities are lower, much larger samples are needed, with this phenomenon being amplified by poorly over-determined (few salient loadings) factors. Indeed sample sizes may have to be much larger than typically recommended (e..g. a 20:1 subjects to variable ratio) when communalities are low and factors are not highly over-determined. Our theoretical framework

along with results of sampling studies, in combination with those of MacCallum et al (1999), provide a more informed view of the issue of sample size in factor analysis. It is now clear that it is not possible to make blanket recommendations regarding this issue without considering other important aspects of design". (p. 636).

References

Green, S. B. (1991) How many subjects does it take to do a regression analysis? *Multivariate Behavioural Research*, 26, 3, 499-510.

Guadagnoli, E., Velicer, W. F. (1988) Relation of sample size to the stability of component patterns. *Psychological Bulletin*, 103, 2, 265-275.

MacCallum, R.C., Widaman, K.F., Zhang, S., Hong, S. (1999) Sample Size in Factor Analysis. *Psychological Methods*, 4, 1, 84-99.

MacCallum, R.C., Widaman, K.F., Preacher, K.J., Hong, S. (2001) Sample Size in Factor Analysis: The role of model error. *Multivariate Behavioural Research*, 34, 6, 611-637.

Nunnally, J.C. (1978) Psychometric Theory, 2nd Edition. New York: McGraw-Hill

Tabachnik, B.G. and Fidell, L.S. (1989) Using Multivariate Statistics 3rd Edition. London: Allyn and Bacon

Tabachnik, B.G. and Fidell, L.S. (2001) Using Multivariate Statistics 4th Edition. London: Allyn and Bacon

Velicer, W.F., Fava, J.L. (1998) Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods*, 3, 2, 231-251