PERSONALITY STUDY AND GROUP BEHAVIOUR, 1981, 1, 1

The Observation to Variable Ratio in Factor Analysis

Paul T. Barrett and Paul Kline

Abstract : *Many investigators have suggested a minimum ratio of observations to variables or an absolute minimum of observations in order to obtain stable factor patterns. In this paper a systematic analysis of the problem is undertaken, contrasting instability due to very low ratios with that due to very low quantities of observations. Two sets of data from the 16 P.F. and EPQ personality questionnaires were randomly split into subsamples, with ratios varying from $1\frac{1}{4}$: 1 to 31 : 1. The results indicated that the observation to variable ratio did not influence factor stability, the important variable being the absolute number of observations.*

Introduction

One of the many problems encountered when conducting a factor analytic research is that of choosing the quantity of observations to be made given a certain number of variables. If too few observations are made upon each variable, then the resulting rotated factor pattern is likely to be unstable. Approaching the limit of a square matrix (in which the number of observations equal the number of variables) has been criticised by many investigators on the grounds of artificially forcing dependencies among correlation coefficients, leading to biased estimates of communalities in common factor analysis. However, this argument is only relevant to matrices of approximately less than 20 variables (Harman, 1976; Nunnally, 1978). Nevertheless, the rank of this square matrix is less than its order, hence it is singular and non-Gramian. It is clearly inappropriate to attempt to factor such a matrix and generalise the results. Simply increasing the number of observations is the obvious solution to this dilemma, but with large variable arrays, the number of observations may become prohibitive in terms of cost, time, and

Request for reprints

availability. Therefore, the problem is simply one of determining the minimum number of observations required to yield replicable, stable, factors.

Many investigators have suggested minimum values expressed either as observation to variable ratios or as absolute numerical limits. Guilford (1954) suggested an absolute minimum of 200 observations when using Pearson correlations. Although providing a different estimate when using tetrachoric coefficients, the use of these coefficients is fundamentally incorrect. This is because the correlation of a variable with a linear combination of variables requires the use of product moment coefficients in the correlation of sums. Gorsuch (1974) has suggested a ratio of about 5 times as many observations as variables. Cattell (1974) has suggested a ratio of 3 to 6 times as many observations as variables. He also indicates an absolute minimum of about 250 observations. Nunnally (1978) has suggested a ratio of 10 times as many observations as variables.

There are two reasons as to why these ratios and numerical values are considered so important to factor stability. Firstly the standard error of the coefficient distribution is determined by the number of observations. The smaller the number, the larger the standard error. The second reason is that based upon consideration of the bias encountered in multiple correlations. Obviously, with a square matrix of observations and variables, all multiple correlations between each variable and all others are 1.00. Thus for initial squared multiple correlation (SMC) starting communality estimation, the values in the main diagonal of a correlation martix will all be biased upwards from the 'true' values. However, as stated above, this is only important (with matrices of less than 20 variables). However, given an orthogonal factor pattern matrix (whether rotated or unrotated), the sum of the squared loadings for each variable across the retained factors equals the SMC of that variable with all the factors (otherwise known as the solution communality). If this value was corrected for bias (Claudy, 1978), the number of variables entering into the correction equation is now the quantity of factors extracted. Thus what might have been a poor observation to variable ratio will become apparently quite respectable, with a corresponding decrease in the value of expected bias. Therefore, following this reasoning the observation to variable ratio per se is seen as irrelevant to the problem of choosing

an appropriate sample size for factor analysis. Rather, the solution is viewed as being dependent upon consideration of the standard error of the initial correlation coefficients between variables and on how representative of a population is the sample of observations (of course, the variables are assumed to be representative of the 'true' factors).

The details below are an empirical test of these arguments and a clarification of how factor loadings behave from small to large sample sizes. The data originates from two samples taken over the Eysenck Personality Questionnaire (EPQ: Eysenck and Eysenck, 1975), and Cattell's 16 P.F. (Cattell, Eber, and Tatsuoka, 1970). Thus while in this case the observations are human volunteer subjects, the arguments above and results below hold for any set of observations submitted to this form of analysis.

Method

Subjects

A sample of 250 male and 241 female students completed form A of the 16 P.F. inventory. The student data were drawn primarily from undergraduates attending Exeter University, with Portsmouth Polytechnic contributing 46 business management trainees. The EPQ responses were those from a Gallup quota sample of 600 English male and 598 English female adults. (The data were kindly loaned to us by Professor H.J. Eysenck). The details and characteristics of this sample are given in Eysenck (1979).

Procedure

For the purposes of this analysis, the 16 P.F. was scale rather than item factored, thus using 16 variables for all analyses. Conversely, the EPQ was item factored, using all 90 variables. Each total sample was split into subsamples, preserving the 1 : 1 ratio of males to females in each subsample. The composition of the subsamples was determined by random sampling of the total data set using uniformly distributed numbers generated by the linear congruential method (Knuth, 1969). A completely different sequence of numbers was generated for each subsample. Table 1 presents the essential details of these samples for both questionnaires.

TABLE 1

Questionnaire Sample Compositions

	16 P.F.		EPQ	
	No. of SUBJ.	RATIO	No. of SUBJ.	RATIO
SAMPLE 1	20	1¼:1	112	1¼:1
SAMPLE 2	32	2:1	180	2:1
SAMPLE 3	48	3:1	270	3:1
SAMPLE 4	96	6:1	540	6:1
SAMPLE 5	192	12:1	810	9:1
SAMPLE 6	288	18:1		
TOTAL SAMPLE	491	31:1	1198	13:1

Note : the ratio is that of subjects to variables

Pearson correlation matrices computed from each subsample and the total sample from each questionnaire were submitted to a principal components analysis. Factor extraction, in the case of the 16 P.F. and EPQ total samples, was determined by consideration of the results from AUTOSCREE (a computer implemented Scree test (Cattell, 1966), the details of which are provided in Barrett and Kline (1980a)), Velicer's (1976) minimum average partial (MAP) test, and Cronbach alpha factor reliabilities (Kaiser and Caffney, 1965). Thus, as the purpose of this investigation is to examine factor stability, the number of factors retained from the total sample were then rotated to a maximum simple structure position using direct oblimin (Jennrich and Sampson, 1966). The δ parameter being swept from —30.0 to 0.6 in steps of 0.1, providing a series of solutions varying from near orthogonality to maximum solution obliquity. For each step the hyperplane count (HC) and sum of squared loadings within the HC bound ± 0.1 (SSL) was noted. The final position and factor pattern (V_p) being given by the maximum HC. The SSL values determined the position within a maximum HC plateau.

From the final factor patterns, Pearson correlations and Tucker congruences were computed between the total sample and all subsample factors for each questionnaire (reflecting and correcting for factor position where necessary). Additionally, an examination of actual loading behaviour was made by considering the total sample as the reference V_{pt} and comparing the subsample V_{ps} with it. For this comparison three groups of loading size were used, from —1.00 to 0.35, from —0.34999 to 0.34999, and from 0.35 to 1.00. Obviously, the larger the absolute value of a loading, the more influential its relevance to the particular factor upon which it loads. This loading error analysis provided information concerning interaction of V_{pt} loading size and V_p errors. Loadings from the entire solution in each case, rather than each factor individually, were compared to the V_{pt} loadings. This effectively ignored factor position effects. With reference to the loading of variable i in V_{pt}, variable i from a subsample V_p was expressed as over or underestimating this value. Thus both the quantity and mean over and underestimates within each group size could be readily computed. For negative loadings, an overestimate was defined as a 'bigger' negative, while for a positive loading an overestimate was defined as a 'bigger' positive.

Results

It was not clear from the analysis of the total sample 16 P.F. data, how many factors should be extracted. While the Velicer test indicated 2 factors, AUTOSCREE gave 7 and the reliability coefficients suggested 4. Although oblique rotation will spread variance across factors, boosting the reliabilities of factors with low eigenvalues, one has to be careful of overfactoring and producing a factor structure with low overall reliability. Thus both 4 and 7 factor solutions were included in the associative analysis, the results from this providing further evidence for the number to be retained. The EPQ results were based upon previous research (Barrett and Kline, 1980b; Barrett and Kline, 1980c) demonstrating a clear 4 factor structure for this test. These factors were found both at the 1st and 2nd order levels, and there was no doubt concerning their number.

Thus all subsamples were factored and rotated according to the extraction decisions above. The associative analysis was then undertaken. Tables 2, 3, and 4 present these results. The top figure in each box is the Pearson correlation, the bottom figure is the congruence coefficient.

TABLE 2

Associative Analysis : 16 P.F. 7 Factor Solution

	TOTAL SAMPLE FACTORS						
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4	FACTOR 5	FACTOR 6	FACTOR 7
SAMPLE 1	0.7961 0.7189	0.5996 0.6058	0.7792 0.7320	0.4789 0.4350	0.2466 0.2269	0.7142 0.6952	0.7393 0.7541
SAMPLE 2	0.9129 0.9113	0.6180 0.6312	0.6854 0.4970	0.8088 0.8045	0.2408 0.2025	0.8537 0.8632	0.8979 0.9027
SAMPLE 3	0.5979 0.5844	0.7721 0.6872	0.8526 0.8854	0.9055 0.9116	0.6224 0.6621	0.7389 0.7616	0.2044 0.1288
SAMPLE 4	0.9610 0.9575	0.9765 0.9732	0.7452 0.6755	0.9510 0.9533	0.3646 0.4184	0.6768 0.7515	0.9566 0.9552
SAMPLE 5	0.9663 0.9659	0.8351 0.7837	0.9539 0.9609	0.9351 0.9367	0.5240 0.5755	0.7391 0.7318	0.9705 0.9720
SAMPLE 6	0.9941 0.9921	0.9913 0.9909	0.9539 0.9653	0.9909 0.9914	0.9403 0.9127	0.8901 0.9129	0.9905 0.9879

Top fig.=the pearson coefficient
Bottom fig.=congruence coefficient

TABLE 3

Associative Analysis : 16 P.F. 4 Factor Solution

	TOTAL SAMPLE FACTORS			
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
SAMPLE 1	0.9233 0.9201	0.9028 0.8985	0.6098 0.6267	0.7170 0.7211
SAMPLE 2	0.8713 0.8719	0.8253 0.8179	0.8422 0.8423	0.8586 0.8604
SAMPLE 3	0.9615 0.9518	0.9534 0.9517	0.8785 0.8828	0.9613 0.9606
SAMPLE 4	0.9615 0.9611	0.9550 0.9390	0.9864 0.9875	0.9458 0.9434
SAMPLE 5	0.9931 0.9932	0.9832 0.9833	0.9888 0.9894	0.9854 0.9814
SAMPLE 6	0.9974 0.9960	0.9933 0.9931	0.9931 0.9937	0.9939 0.9933

Top fig.=the pearson coefficient
Bottom fig.=congruence coefficient

TABLE 4

Associative Analysis : EPQ 4 Factor Solution

	TOTAL SAMPLE FACTORS			
	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
SAMPLE 1	0.9299 0.9420	0.9380 0.9490	0.9062 0.9246	0.8884 0.8982
SAMPLE 2	0.9518 0.9607	0.9581 0.9660	0.9590 0.9666	0.9298 0.9319
SAMPLE 3	0.9794 0.9826	0.9688 0.9750	0.9543 0.9630	0.9728 0.9776
SAMPLE 4	0.9870 0.9898	0.9869 0.9884	0.9797 0.9818	0.9794 0.9812
SAMPLE 5	0.9965 0.9973	0.9952 0.9961	0.9928 0.9939	0.9937 0.9946

Top fig.=the pearson coefficient

Bottom fig.=congruence coefficient

Accepting that a value of >0.75 for both coefficients indicates sufficient similarity between factors, it is clear that the 7 factor 16 P.F. solution is unsatisfactory. This is viewed as a direct result of overfactoring. A large quantity of error variance is being distributed across the solutions leading to factor instability. The alpha coefficients for the rotated 6th and 7th factors from the total sample analysis are very low indeed : $\alpha_6 = 0.27$, $\alpha_7 = 0.09$. Therefore, this particular solution was discarded in place of the 4 factor representation. Since both this and the EPQ solution demonstrated sufficient factor stability across the subsamples, the loading error analysis was undertaken so as to indicate the amount of error entering into factor loading estimates as compared with the total sample loadings. Tables 5 and 6 present the results of this analysis.

TABLE 5

Loading Error Analysis : 16 P.F. 4 Factor Solution

		SAMPLE 1		SAMPLE 2		SAMPLE 3		SAMPLE 4		SAMPLE 5		SAMPLE 6	
GROUPSIZE		%	MEAN	%	MEAN	%	MEAN	%	MEAN	%	MEAN	%	MEAN
-1.00 to -0.35	+	66.7	0.133	44.4	0.189	55.6	0.088	44.4	0.049	22.2	0.021	55.6	0.031
(9)	-	33.3	0.088	55.6	0.184	44.4	0.103	55.6	0.107	77.8	0.045	44.4	0.017
-0.34999 to +0.34999	+	32.6	0.190	45.7	0.206	50.0	0.127	47.8	0.093	43.5	0.047	50.0	0.037
(46)	-	67.4	0.199	54.3	0.149	50.0	0.072	52.2	0.083	56.5	0.048	50.0	0.037
+0.35 to +1.00	+	44.4	0.159	44.4	0.127	33.3	0.056	55.6	0.087	66.7	0.060	55.6	0.026
(9)	-	55.6	0.351	55.6	0.197	66.7	0.130	44.4	0.085	33.3	0.030	44.4	0.031

\+ = overestimate

— = underestimate

% = percentage over and underestimated within a particular groupsize

MEAN = mean over and underestimates

( ) = figures in brackets indicate the number of total sample loadings within the groupsize

TABLE 6

Loading Error Analysis : EPQ 4 Factor Solution

		SAMPLE 1		SAMPLE 2		SAMPLE 3		SAMPLE 4		SAMPLE 5	
GROUPSIZE		%	MEAN	%	MEAN	%	MEAN	%	MEAN	%	MEAN
-1.00 to -0.35	+	40.0	0.065	45.0	0.061	65	0.034	35	0.028	50	0.016
(20)	-	60.0	0.085	55.0	0.075	35	0.026	65	0.045	50	0.022
-0.34999 to +0.34999	+	50.3	0.071	50.3	0.063	55.6	0.048	41.9	0.034	45.1	0.018
(284)	-	49.7	0.086	49.7	0.059	44.4	0.049	58.1	0.037	54.9	0.019
+0.35 to +1.00	+	44.6	0.067	50.0	0.057	48.2	0.042	33.9	0.021	44.6	0.014
(56)	-	55.4	0.070	50.0	0.057	51.8	0.039	66.1	0.029	55.4	0.020

\+ = overestimate

— = underestimate

% = percentage over and underestimated within a particular groupsize

MEAN = mean over and underestimates

( ) = figures in brackets indicate the number of total sample loadings within the groupsize

Discussion

From the results in Tables 3 and 5, factor stability in the 16 P.F. 4 factor solution is demonstrated in samples 3, 4, 5 and 6. For sample 3 (N=48) the mean errors of over and underestimates are all near the value of 0.1 with all but factor 3 measures of association >0.95. The observation (subjects) to variable ratio in this sample is 3:1. From Tables 4 and 6, the EPQ factors are stable across all samples. For sample 1 (N=112) the mean errors of over and underestimates are <0.09 with association coefficients >0.9 for all but factor 4 (0.8884 and 0.8982). The ratio being $1\frac{1}{4}$: 1. Noticeably in both solutions, the number of over and underestimates within each groupsize is relatively equal, there are no gross trends within the data. Additionally, there are no specific loading size effects. That is, the size of a loading does not determine its stability, numerically low value loadings are not more unstable than those of higher value.

Thus from the above, it would appear that observation to variable ratio has no effect on factor stability. Rather it is the number of observations that is the crucial feature. On the basis of the results presented here, the minimum quantity of observations required to yield a clear, recognisable factor pattern is 50. However, both sets of data contained good 'strong' variables yielding exceedingly clear factor structures from the total samples. Also, the factorial constructs being manipulated in this analysis are assumed (in trait theory) to be general pervasive influences in all individuals. Thus virtually any sample of individuals would suffice to yield the underlying structures. The argument here is that statistical error is relatively minimal compared with the errors to be found from bad sampling of target populations.

In addition, observing the instability of the factors in the 7 factor 16 P.F. solution (Table 2), it is clear that factor extraction prior to rotation is crucial in determining a 'correct' factor structure. For these reasons, small sample factoring can only be carried out, practically, on variables which have a known factor structure, with the purpose of replicating the supposed structure and maintaining a constant check on the behaviour of the variables. If attempting to small sample factor a set of untried variables, as in a pilot study, it is crucial to concentrate maximum efforts on determining the number of factors to be extracted. Several solutions will no doubt have to be rotated and the most satisfactory in terms of factor

reliabilities and validities (Cattell and Tsujioka, 1964) chosen. However, it is essential that at least 2 small samples are taken in order to cross validate the structures.

In conclusion, the theoretical argument stated in the introduction concerning the expected lack of SMC bias, reflected in individual variable loadings, has been empirically validated implying that the subject to variable rates (given that it is greater than 1 : 1) is not important. Factor stability, given known loading behaviour and adequate target population sampling, is simply a function of the accuracy of initial correlation estimates. If the standard errors are large, any resulting factors structure is likely to be masked by excessive statistical errors of measurement.

References

Barrett, P. T. and Kline, P. Factor extraction : an examination of three methods. 1980, Submitted for publication (a).

Barrett, P. T. and Kline, P. Personality factors in the Eysenck Personality Questionnaire. *Personality and Individual Differences*, 1980, In Press (b).

Barrett, P. T. and Kline, P. The Itemetric properties of the Eysenck Personality Questionnaire : a reply to Helmes. 1980, Submitted for publication (c).

Cattell, R. B. The Scree test for the number of factors. *Multivariate Behavioural Research*, 1966, 1, 245-276.

Cattell, R. B. *The Scientific Use of Factor Analysis in Behavioural and Life Sciences*. New York : Plenum Press, 1978.

Cattell, R. B., Eber, H.W., and Tatsuoka, M.M. *Handbook for the Sixteen Personality Factor Questionnaire*. Illinois : Institute for Personality and Ability Testing, 1970.

Cattell, R. B. and Tsujioka, B. The Importance of factor trueness and validity, versus homogeneity and orthogonality, in test scales. *Educational and Psychological Measurement*, 1964, 24, 3-30.

Claudy, J. G. Multiple regression and validity estimation in one sample. *Applied Psychological Measurement*, 1978, 2, 595-607.

Eysenk, H.J. Personality factors in a random sample of the population. *Psychological Reports*, 1979, 44, 1023-1027.

Eysenck, H.J. and Eysenck, S.B.G. *Manual of the Eysenck Personality Questionnaire* London : Hodder and Stoughton, 1976.

Gorsuch, R.L. *Factor Analysis*. Philadelphia ; Saunders, 1974.

Guilford, J.P. *Psychometric Methods (2nd Edit.)*. New York : McGraw-Hill, 1954.

Harman, H.H. *Modern Factor Analysis (3rd Edit.)*. Chicago : University of Chicago Press, 1976.

Jennrich, R. I. and Sampson, P. F. Rotation for simple loadings. *Psychometrika*, 1966, 31, 313-323.

Kaiser, H. F. and Caffrey, J. Alpha factor analysis. *Psychometrika*, 1965, 30, 1-14.

Knuth, D. E. *The Art of Computer Programming*, Addison-Wesley, 1969.

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

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