Psychological Bulletin
1988, Vol. 103, No. 2, 265-275

Copyright 1988 by the American Psychological Association, Inc.
0033-2909/88/$00.75

Relation of Sample Size to the Stability of Component Patterns

Edward Guadagnoli and Wayne F. Velicer
University of Rhode Island

A variety of rules have been suggested for determining the sample size required to produce a stable solution when performing a factor or component analysis. The most popular rules suggest that sample size be determined as a function of the number of variables. These rules, however, lack both empirical support and a theoretical rationale. We used a Monte Carlo procedure to systematically vary sample size, number of variables, number of components, and component saturation (i.e., the magnitude of the correlation between the observed variables and the components) in order to examine the conditions under which a sample component pattern becomes stable relative to the population pattern. We compared patterns by means of a single summary statistic, g^2, and by means of direct pattern comparisons using the kappa statistic. Results indicated that, contrary to the popular rules, sample size as a function of the number of variables was not an important factor in determining stability. Component saturation and absolute sample size were the most important factors. To a lesser degree, the number of variables per component was also important, with more variables per component producing more stable results.

Factor analysis or component analysis is typically used by the researcher who wishes to reduce a set of observed variables, p, to a new, smaller set of variables. This smaller set of new variables (m), labeled *factors* or *components,* depending on the method used, preserves most of the information present in the original set of variables and is a more parsimonious representation. The purpose of an analysis may be the replacement of the p scores with m factor or component scores or the interpretation of the $p \times m$ pattern of loadings, that is, correlations between the p observed variables and the m factors or components. The latter is intended to facilitate the understanding of the relations that exist between the observed variables.

A major issue involves determining the number of independent observations (N) required to obtain a sample pattern that is stable and approximates the population pattern. Researchers and textbook authors typically recommend that the necessary sample size be determined as a function of the size of p involved in the research problem. (See Baggaley, 1982; Brislin, Lonner, & Thorndike, 1974; Cattell, 1952, 1978; Gorsuch, 1983; Hair, Anderson, Tatham, & Grablowsky, 1979; Kunce, Cook, & Miller, 1975; Lindeman, Merenda, & Gold, 1980; Marascuilo & Leven, 1983; Nunnally, 1978; for suggested N-to-p ratios, which vary from 2:1 to 20:1.) A minimum N of 100 to 200 observations is also often recommended (Comrey, 1973, 1978; Gorsuch, 1983; Guilford, 1954; Hair et al., 1979; Lindeman et al., 1980; Loo, 1983). The rules relating N to p seem to be based on the shrinkage concept developed in multiple regression. 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. Another rule suggests that N be determined as a function of the number of expected factors (Cattell, 1978). The most familiar advice given the researcher, however, is to obtain the maximum sample size possible (Guertin & Bailey, 1970; Humphreys, Ilgen, McGrath, & Montanelli, 1969; Press, 1972; Rummell, 1970).

A variety of rules have been suggested in the literature, and none of them is empirically based. Rather, rules appear to be generated as a function of the experience of an author, an unstated set of beliefs, or communication from some uncited expert source.

In a limited number of studies, researchers have empirically tested the relation between sample size and the stability of the sample solution. Aleamoni (1973) used a real-data matrix involving a sample of 2,322 as the population and generated subsamples of 17, 25, 100, 400, and 1,600. For each subsample, he performed a principal-components analysis on a 15-by-15 correlation matrix. Aleamoni found that the variance accounted for increased as sample size decreased, indicating that smaller samples may include more error variance. He found that component similarity also decreased with a reduction in sample size. Aleamoni (1973) concluded, "If we want to use sample factor structures as a basis for generalizing to their corresponding population factor structures, drawing random samples of $N = 400$ is adequate for generalizing to a population of $N = 2,322$" (p. 269). Barrett and Kline (1981), using real data from the Sixteen Personality Factor Questionnaire, created subsamples of various sizes for both instruments and concluded that sample size as a function of the number of variables did not influence pattern stability. An N of 50 was the minimum needed to reproduce the pattern. Arrindell and van der Ende (1985), using real data from the Fear Survey Schedule–III and the Fear Questionnaire, reached the same conclusion about the

An earlier version of this article was presented at the Psychometric Society meeting in Nashville, Tennessee, in June 1985. This work was partially supported by Grant CA 27821 from the National Cancer Institute.

Edward Guadagnoli is now at the Center for Health Care Research, Brown University, Providence, Rhode Island 02912.

Correspondence concerning this article should be addressed to Wayne F. Velicer, Department of Psychology, University of Rhode Island, Kingston, Rhode Island 02881.

utility of an observations-to-variables ratio. They also found no support for an absolute minimum, but the smallest samples were 78 and 100.

Velicer, Peacock, and Jackson (1982), in a simulation study, compared the solutions obtained from three types of factor analysis procedures (principal-components analysis, image-component analysis, and maximum likelihood factor analysis) to determine under what conditions the methods produce different patterns. A comparison of sample factor and component patterns to their respective population factor and component patterns suggested that "with only moderate sample sizes (N = 144), the fit of the pattern to the population target was quite good" (Velicer et al., 1982, p. 386).

All previous empirical investigations provide results that are limited in scope, focusing on one data set or one value of p; their results, however, provide a framework for future research.

The purpose of the present study was to determine the conditions necessary to produce a stable solution with respect to a population pattern. We used a principal-components procedure (Hotelling, 1933) to produce population and sample component patterns from computer-generated correlation matrices. Although factor analysis procedures are commonly recommended, Velicer (1974, 1976, 1977), Velicer and Fava (1987), and Velicer et al. (1982) have demonstrated that principal-components solutions differ little from the solutions generated from factor analysis methods. Additionally, serious theoretical problems exist with the factor analysis model (Steiger & Schönemann, 1978). Component analysis does not suffer from some of the convergence problems, boundary cases, and computational limitations (Driel, 1978; Velicer & Fava, 1987) that factor analysis does, permitting assessment of a wider range of situations (Jackson & Chan, 1980; Velicer & Jackson, 1987). Further, Glass and Taylor (1966), following a survey of educational research journals, reported that component analysis was the most frequently performed analysis. Pruzek and Rabinowitz (1981), over a decade later, reported that this trend had not only continued but had increased.

To evaluate the conditions under which stable component patterns are produced, we manipulated several variables. We selected P, m, and N as potentially important factors with respect to the rules already discussed. In addition to these factors, component saturation (a_{ij}), that is, the magnitude of component loadings, was also varied. Correlation matrices were computer generated so that direct control over the variables manipulated was possible.

Method

With the design of this study, we attempted to sample conditions that are often encountered by applied researchers. The situations generated, however, were typically simpler than real-world conditions. The population component patterns included only variables that loaded on a single component. Each component was defined by an equal number of variables. All nonzero loadings were equal. Table 1 contains an example of the type of pattern included in this study. These conditions represent a relatively clean, or idealized, condition. However, an examination of these ideal patterns represents a necessary first step in the evaluation of this problem. Computer generation of correlation matrices allowed manipulation and comparison of known conditions but was less realistic than the use of real data sets. Tucker, Koopman, and Linn (1969) discussed problems associated with the use of simulated data versus real data.

Table 1
Example of Population Pattern Matrix (p = 12, m = 3, a_{ij} = .60)

Observed variable (p)	Component (m)		
	1	2	3
1	.60	.00	.00
2	.60	.00	.00
3	.60	.00	.00
4	.60	.00	.00
5	.00	.60	.00
6	.00	.60	.00
7	.00	.60	.00
8	.00	.60	.00
9	.00	.00	.60
10	.00	.00	.60
11	.00	.00	.60
12	.00	.00	.60

First, we performed a principal-components analysis of both population and sample correlation matrices. Second, we compared the resulting sample component patterns with their respective population component patterns. Comparison methods involved both the least squares difference between corresponding patterns and a direct pattern comparison approach. The former involved a summary statistic, g^2, which measured the average squared difference between comparable loadings of the sample and population patterns. The latter comparison involved examining agreement with respect to the identification of salient and nonsalient component loadings.

Factors Manipulated

We examined the similarity of sample component patterns to population patterns across seven levels of N, four levels of p, three levels of a_{ij}, and three levels of m. We chose sample sizes of 50, 100, 150, 200, 300, 500, and 1,000 not only to represent a range of small to large samples but also to examine precisely the sample-size range suggested by previous research. Number of variables ranged from 36 to 144. Intermediate levels included 72 and 108. We eliminated from the design cases in which $N < p$. Nunnally (1978) summarized the arguments against violating this minimum. We selected 3, 6, and 9 for the number of components. Values of m much larger than 9 are not desired by applied researchers because interpretation may become a problem at this point. In the interest of practicality, however, we did not examine the cases in which m = 3 and p = 108 or p = 144 because the number of variables per component (48 and 36, respectively) represent unusual situations. For cases in which p = 108 and p = 144, we substituted the value m = 18. We used three levels of a_{ij}. With respect to principal-components analysis, component loadings of .30 or .40 are usually regarded as salient to that particular component, whereas loadings below the cutoff are ignored. Loadings used in this design include (a) a typical lower limit, a_{ij} = .40; (b) a moderate level, a_{ij} = .60; and (c) a very well-defined value, a_{ij} = .80.

Data Generation

The conditions described represent a design in which 252 combinations ($7 \times 4 \times 3 \times 3$) are possible. Of these, only 207 combinations were examined, however, given the limitation $N > p$.

We constructed population matrices for each possible p, m, and a_{ij} combination. We generated population correlation matrices following a procedure used by Zwick and Velicer (1982, 1986), Velicer and Fava (1987), and Velicer et al. (1982). A $p \times m$ population pattern matrix (A)

was generated with respect to every possible p, m, and a_{ij} combination defined in the design. Postmultiplying **A** by its transpose (**A'**) generated a $(p \times p)$ matrix, $\mathbf{R^*(AA' = R^*)}$. We obtained the population correlation matrix (**R**) by replacing the elements in the diagonal of the **R*** matrix with unities. Table 1 contains an example of an **A** matrix for the case in which $p = 12$, $m = 3$, and $a_{ij} = .60$. We used a computer program developed by Montanelli (1975) to generate five sample correlation matrices from each **R** for every level of sample size in the study. We performed a principal-components analysis to obtain population and sample component patterns from the respective correlation matrices.

Pattern Comparison

We used a summary statistic, g^2, to compare sample component patterns with population patterns. This statistic is based on the usual least squares criterion. First, a difference matrix (**E**) is calculated from the sample component pattern (**A**) and the population pattern (**A***), where

$$\mathbf{E} = \mathbf{A} - \mathbf{A}^*. \quad (1)$$

The summary statistic, g^2, is then defined as

$$g^2 = \text{trace}\,(\mathbf{E'E})/pm, \quad (2)$$

which can be interpreted as the average (squared) difference between comparable loadings of corresponding sample and population component patterns (Velicer, 1977; Velicer & Fava, 1987; Velicer et al., 1982). We made comparisons following a varimax rotation (Kaiser, 1958) of the population and sample component patterns. To facilitate comparison between patterns, we generated a permutation matrix (Velicer, 1974, 1976, 1977). This permutation matrix allowed a one-to-one component match with the population pattern by permuting the columns of the sample component pattern. To simplify interpretation, we selected .01 as the maximum g^2 value for describing an acceptable fit between sample and population component patterns. Values below this cutoff imply that on the average, the difference between comparable loadings of the population and sample component pattern occurs only in the second decimal place. We calculated the average g^2 over the five samples generated for each condition.

The g^2 statistic has the advantage of being a single scalar value with a direct operational interpretation. It also possesses several potential drawbacks. First, it will be affected by shrinkage. In principal-components analysis, shrinkage takes the form of inflating the initial eigenvalues and decreasing the value of the later eigenvalues (Bobko & Schemmer, 1984). This could result in inflated loadings with smaller sample sizes and correspondingly larger values of g^2, even though the pattern correctly reproduces all the essential positions or salience information. Second, the g^2 statistic represents an average and may be affected by one or two extreme values or one noncorresponding column. This can produce the same result as a large number of small over- or underestimates. The former would result in interpretation errors, whereas the latter would not. This second problem is the result of the lack of configural, or position, information in the g^2 statistic. For these reasons, we also compared population and sample component patterns by using a method previously used by Velicer et al. (1982). This method involves first determining those variables that are considered salient to a component. Following common practice, we determined a variable to be salient if the component loading was greater than .40 (Velicer et al., 1982). We then compared salient loadings present in the sample pattern with salient loadings in the population pattern. We generated decision tables that described the results of the comparison with respect to hits and misses. Two types of error (misses) were possible: Type I error, in which a variable is judged salient when it is not salient, and Type II error, in which a variable is not judged to be salient when it actually is salient. *Hits* referred to the correct identification of salient and nonsalient variables between patterns. We constructed decision tables following comparisons of the varimax-rotated sample pattern with the corresponding population solution.

A great variety of agreement statistics (Fleiss, 1981) may be calculated from the type of decision table described. We used the kappa statistic, a measure of agreement, in the present study. In addition to providing a correction for chance expected agreement, kappa is the appropriate agreement measure to use under conditions that involve comparison with a standard or correct set of responses (Light, 1971). The population component pattern state of affairs (component loading salient or not salient) represents the standard with which the sample component pattern is compared. Complete agreement between the two patterns is defined by a kappa value of 1.00. Kappa $\geq .00$ signifies agreement greater than or equal to chance level; values below .00 represent agreement below chance level. Landis and Koch (1977) provided guidelines for interpreting kappa. Kappa values greater than .75 represent excellent agreement beyond the chance level, values between .40 and .75 are indicators of fair to good agreement beyond the chance level, and values below .40 represent poor agreement.

We calculated three summary statistics—kappa, Type I error, and Type II error—from the pooled decision tables resulting from the five samples generated for each combination of the variables manipulated. Type I error was scaled by the number of nonsalient loadings that should have occurred in a particular pattern. We used division by the number of nonsalient loadings, $p(m - 1)$, to allow comparison across levels of p used in the study. The resulting value multiplied by 100 represents the average percentage of possible Type I error classifications. Similarly, Type II error was scaled by the number of actual salient loadings (p) that should have been present in the pattern and is presented as the average percentage of possible Type II error classifications.

Results

We derived five sample correlation matrices, generated for each level of sample size, from each of the 36 population correlation matrices (defined by a combination of p, m, and a_{ij} levels). In several instances, a subset or all of the resulting sample component patterns did not possess a structure defined well enough for a one-to-one component match with the population component structure to be attained. That is, the permutation matrix used to match the sample pattern with the population pattern could not be generated. In cases in which five matches were not accomplished, we increased the number of matrices generated until five matches were attained. Matching problems occurred only for the low-saturation condition (.40). The conditions under which more matrices were generated are presented in Table 2. Note that we dropped one condition ($p = 108$, $a_{ij} = .40$, $m = 18$, $N = 150$) from the study because we could find only one match after 30 samples had been generated.

g^2 *Statistic Comparisons*

Values of g^2 increased as component saturation decreased. At $a_{ij} = .60$ and $a_{ij} = .80$, the g^2 value was below the criterion level used to describe a good fit ($g^2 \leq .01$) for all but the smallest sample sizes ($N = 50$ for $a_{ij} = .80$, $N = 150$ for $a_{ij} = .60$). At these saturation levels, performance generally did not vary as a result of the value of p or m. Differences in g^2 that did result as a function of p or m occurred in the third decimal place. Tables 3 and 4 contain detailed results for these conditions.

Performance at $a_{ij} = .40$ was not as consistent as performance at higher saturation levels. Table 2 contains detailed results for this condition. That is, the effects of p, m, and N on g^2 were observed. The relation between p and g^2 was opposite that ex-

Table 2
Average g^2, Type I Error, Type II Error, and Kappa for 69 Patterns With .40 Loadings

Pattern	*p*	*m*	*N*	g^2	Type I error	Type II error	Kappa
1	36	3	50	.0555	9	40	.54
2	36	3	100	.0153	1	23	.81
3	36	3	150	.0110	0	25	.80
4	36	3	200	.0079	0	17	.87
5	36	3	300	.0049	0	8	.94
6	36	3	500	.0029	0	4	.97
7	36	3	1,000	.0015	0	2	.99
8	72	3	100	.0124	0	32	.74
9	72	3	150	.0075	0	27	.79
10	72	3	200	.0053	0	28	.78
11	72	3	300	.0037	0	27	.79
12	72	3	500	.0022	0	16	.87
13	72	3	1,000	.0012	0	9	.93
14[a]	36	6	50	.0499	8	42	.52
15[b]	36	6	100	.0284	3	30	.72
16[b]	36	6	150	.0226	1	26	.79
17	36	6	200	.0117	0	11	.93
18	36	6	300	.0087	0	6	.96
19	36	6	500	.0043	0	0	1.00
20	36	6	1,000	.0024	0	0	1.00
21	72	6	100	.0153	0	29	.79
22	72	6	150	.0106	0	25	.83
23	72	6	200	.0079	0	18	.88
24	72	6	300	.0051	0	17	.89
25	72	6	500	.0032	0	6	.96
26	72	6	1,000	.0015	0	1	.99
27	108	6	150	.0089	0	26	.82
28	108	6	200	.0068	0	25	.83
29	108	6	300	.0042	0	20	.87
30	108	6	500	.0025	0	12	.93
31	108	6	1,000	.0013	0	4	.97
32	144	6	150	.0080	0	33	.77
33	144	6	200	.0062	0	32	.78
34	144	6	300	.0040	0	25	.83
35	144	6	500	.0024	0	22	.85
36	144	6	1,000	.0012	0	12	.93
37[a]	36	9	50	.0483	6	42	.49
38[a]	36	9	100	.0364	4	37	.61
39[b]	36	9	150	.0253	3	26	.73
40	36	9	200	.0204	1	19	.82
41[b]	36	9	300	.0138	0	11	.91
42	36	9	500	.0077	0	4	.97
43	36	9	1,000	.0032	0	0	1.00
44[a]	72	9	100	.0248	2	41	.65
45[b]	72	9	150	.0159	0	32	.77
46	72	9	200	.0114	0	24	.85
47	72	9	300	.0066	0	9	.94
48	72	9	500	.0039	0	3	.99
49	72	9	1,000	.0021	0	1	.99
50[b]	108	9	150	.0110	0	28	.82
51	108	9	200	.0096	0	25	.84
52	108	9	300	.0058	0	19	.88
53	108	9	500	.0031	0	7	.96
54	108	9	1,000	.0016	0	1	.99
55	144	9	150	.0101	0	35	.77
56	144	9	200	.0072	0	26	.84
57	144	9	300	.0048	0	19	.88
58	144	9	500	.0027	0	13	.92
59	144	9	1,000	.0015	0	4	.98

Table 2 (*continued*)

Pattern	p	m	N	g^2	Type I error	Type II error	Kappa
60[c]	108	18	150				
61[a]	108	18	200	.0131	0	32	.78
62[b]	108	18	300	.0097	0	24	.85
63[b]	108	18	500	.0059	0	9	.95
64	108	18	1,000	.0026	0	0	.99
65[a]	144	18	150	.0139	0	42	.70
66[a]	144	18	200	.0119	0	43	.71
67	144	18	300	.0078	0	25	.85
68	144	18	500	.0045	0	11	.94
69	144	18	1,000	.0020	0	1	.99

[a] Condition required 10 or more additional correlational matrices to be generated. [b] Condition required less than 10 additional samples to be generated. [c] Condition eliminated from the design.

pected. The larger the variable sets, the smaller the sample size necessary to attain a cutoff value of $g^2 = .01$ at smaller sample sizes. In general, larger variable sets ($p > 36$) yielded homogeneous values across all conditions. As sample size increased, however, the performance of g^2 at $p = 36$ corresponded more closely to that of the larger variable sets.

The effect of m on g^2 was also evident at $a_{ij} = .40$. Greater values of m (i.e., less well-defined components in terms of p/m) resulted in increased g^2 values for any combination of p and N. For all levels of p, the rate of decrease in g^2 as sample size increased was slower as m increased. That is, attaining the cutoff level of $g^2 = .01$ required larger sample sizes as m increased for any level of p.

Whereas differences in performance among levels of p and m were observed at this saturation level, the cutoff value established was ultimately attained and surpassed under all conditions when a sufficient sample size was reached. As already discussed, the sample size required to reach the cutoff level was higher for smaller p and for increased m. The largest sample size required to meet the $g^2 = .01$ criterion was 450 for $p = 36$ and $m = 9$. Across all levels of m, the sample size required to meet the cutoff for the remaining p levels ($p > 36$) was smaller ($N = 150$ to $N = 300$). Figure 1 contains three curves, one for each value of a_{ij}, which illustrate the relation between N and the square root of g^2. We use the square root of g^2 in this illustration because it is more directly interpretable.

Decision Table Comparisons

Kappa, Type I error, and Type II error results were similar to those obtained with respect to the g^2 statistic. For kappa, perfect performance (kappa = 1.00) resulted across all levels of p, m, and N at the $a_{ij} = .80$ saturation level (see Table 4). This performance level was also attained when $N = 100$ or greater for all conditions when $a_{ij} = .60$ (see Table 3). At $N = 50$, a kappa of .84 was the lowest attained for any combination of p and m at this saturation level.

As with g^2, the effects of m, p, and N became most evident at $a_{ij} = .40$ (see Table 2). An increase in sample size, no matter what level of p or m, resulted in higher kappa values. The criterion value of .75 was attained at sample sizes beyond 200 for any combination of p and m. Lower levels of p (36 and 72) displayed poorer performance than did higher levels (108 and 144) at low sample sizes ($N = 50$ to $N = 250$). At the larger sample sizes ($N = 300$ to $N = 1{,}000$), even the smaller variable sets possessed good kappa values. An increase in m resulted in a decrease in kappa.

Type I error, the incorrect identification of a nonsalient component loading, was relatively rare under any condition. At $a_{ij} = .80$, Type I error was nonexistent. Type I error at $a_{ij} = .60$ did not occur beyond $N = 100$ for any combination of p and m. Type I error at this saturation level did not surpass 3%. Type I error did become more frequent at $a_{ij} = .40$ (see Table 2). The percentage did not, however, exceed 9%. An increase in N resulted in lower Type I error rates. Smaller variable sets resulted in higher percentages. The difference in percentage of error between levels of p was enhanced with an increase in m. Overall, Type I error was relatively rare.

Compared with Type I error, Type II error, which is the failure to identify salient component loadings, was more frequent. As with the other comparison statistics, Type II errors were observed most frequently at the .40 saturation level. At $a_{ij} = .80$, no Type II errors occurred. At sample sizes of 100 and beyond, Type II errors did not occur at the .60 saturation level. (An exception was $p = 108, 144$; $m = 18$, $N = 150$; Type II error = 1%.) For $N = 50$, Type II error never exceeded 11% ($m = 9$). The $a_{ij} = .40$ level of component saturation provided observable effects of p, m, and N on this statistic. An increase in N within any p and m combination resulted in decreased Type II error percentages. Smaller variable sets generally resulted in higher Type II error percentages. Errors increased as the number of components increased. As m increased, performance as a function of p became more similar, particularly at larger sample sizes (beyond $N = 200$). At low sample sizes ($N = 50$ to $N = 200$), for any level of p, nearly one third of all component loadings that should have been identified as salient were misidentified.

The Prediction of Y

The purpose of this analysis was to determine an approximate rule to assist the applied researcher in determining when the sample pattern will provide an adequate estimate of the population pattern. *Adequate* is assumed to be defined within the context of the particular study. To accomplish this, we used the factors involved in this study (N, a_{ij}, p, and m) as predictors and

Table 3
Average g^2, Type I Error, Type II Error, and Kappa for 69 Patterns With .60 Loadings

Pattern	p	m	N	g^2	Type I error	Type II error	Kappa
1	36	3	50	.0176	0	4	.96
2	36	3	100	.0078	0	1	.99
3	36	3	150	.0049	0	0	1.00
4	36	3	200	.0037	0	0	1.00
5	36	3	300	.0024	0	0	1.00
6	36	3	500	.0014	0	0	1.00
7	36	3	1,000	.0007	0	0	1.00
8	72	3	100	.0069	0	0	1.00
9	72	3	150	.0015	0	0	1.00
10	72	3	200	.0034	0	0	1.00
11	72	3	300	.0023	0	0	1.00
12	72	3	500	.0013	0	0	1.00
13	72	3	1,000	.0007	0	0	1.00
14	36	6	50	.0236	1	7	.92
15	36	6	100	.0089	0	0	1.00
16	36	6	150	.0061	0	0	1.00
17	36	6	200	.0044	0	0	1.00
18	36	6	300	.0032	0	0	1.00
19	36	6	500	.0018	0	0	1.00
20	36	6	1,000	.0009	0	0	1.00
21	72	6	100	.0084	0	1	1.00
22	72	6	150	.0055	0	0	1.00
23	72	6	200	.0041	0	0	1.00
24	72	6	300	.0024	0	0	1.00
25	72	6	500	.0016	0	0	1.00
26	72	6	1,000	.0008	0	0	1.00
27	108	6	150	.0050	0	0	1.00
28	108	6	200	.0038	0	0	1.00
29	108	6	300	.0025	0	0	1.00
30	108	6	500	.0015	0	0	1.00
31	108	6	1,000	.0008	0	0	1.00
32	144	6	150	.0052	0	0	1.00
33	144	6	200	.0037	0	0	1.00
34	144	6	300	.0025	0	0	1.00
35	144	6	500	.0015	0	0	1.00
36	144	6	1,000	.0008	0	0	1.00
37	36	9	50	.0277	2	11	.84
38	36	9	100	.0117	0	0	.99
39	36	9	150	.0077	0	0	1.00
40	36	9	200	.0053	0	0	1.00
41	36	9	300	.0034	0	0	1.00
42	36	9	500	.0021	0	0	1.00
43	36	9	1,000	.0010	0	0	1.00
44	72	9	100	.0096	0	1	.99
45	72	9	150	.0061	0	0	1.00
46	72	9	200	.0047	0	0	.99
47	72	9	300	.0030	0	0	1.00
48	72	9	500	.0018	0	0	1.00
49	72	9	1,000	.0009	0	0	1.00
50	108	9	150	.0056	0	0	.99
51	108	9	200	.0042	0	0	1.00
52	108	9	300	.0028	0	0	1.00
53	108	9	500	.0016	0	0	1.00
54	108	9	1,000	.0008	0	0	1.00
55	144	9	150	.0056	0	0	.99
56	144	9	200	.0040	0	0	1.00
57	144	9	300	.0026	0	0	1.00
58	144	9	500	.0016	0	0	1.00
59	144	9	1,000	.0008	0	0	1.00

Table 3 (*continued*)

Pattern	p	m	N	g^2	Type I error	Type II error	Kappa
60	108	18	150	.0077	0	1	.99
61	108	18	200	.0052	0	0	1.00
62	108	18	300	.0035	0	0	1.00
63	108	18	500	.0020	0	0	1.00
64	108	18	1,000	.0010	0	0	1.00
65	144	18	150	.0064	0	0	.99
66	144	18	200	.0049	0	0	1.00
67	144	18	300	.0031	0	0	1.00
68	144	18	500	.0018	0	0	1.00
69	144	18	1,000	.0009	0	0	1.00

used the summary statistic, g^2, as the criterion. We examined a large number of models and selected the final best solution on the basis of three criteria: (a) the size of the multiple correlation, (b) the simplicity of the solution, and (c) the meaningfulness of the solution.

The models investigated included a variety of transformations and higher order predictors, in addition to the four predictors already mentioned. Higher order terms involved both powers of the predictors (i.e., X^2) and cross products of predictors (e.g., X_iX_j). We paid special attention to those terms that have been advocated previously, such as the ratios N/p and p/m. Transformations that were the most meaningful were also the most successful. The sample-size variable was transformed to the reciprocal of the square root of N, a transformation suggested by the standard error of a correlation coefficient. We used the square root of g^2 as the criterion and relabeled it Y. These transformations provided both a better statistical fit and a more direct interpretation. These two transformations permitted the fitting of a linear equation to the previously nonlinear relation (see Figure 1).

The best fitting multiple regression equation involved nine predictors (the four simple predictors and five higher order predictors) and resulted in a multiple correlation of .93. However, a two-predictor equation involving rounded weights produced almost as good a fit (R = .92) and was judged preferable on the basis of both the meaningfulness and the simplicity criteria. This equation was

$$Y = 1.10(X_1) - .12(X_2) + .066, \quad (3)$$

where Y is the average distance between a population loading and a sample loading; X_1 is the reciprocal of the square root of N, approximately the standard error of a correlation coefficient; and X_2 is the average loading on a salient variable.

The following example illustrates how this equation might be used: A researcher estimates that the average loading for those variables that will be salient for a component is .60. For a sample size of 100, Y is estimated to be .104. For a sample size of 200, Y is estimated to be .072. For a sample size of 400, Y is estimated to be .049. The researcher could use these estimates to determine which sample size will produce a sample pattern that is adequate for the research purposes.

The most difficult predictor to estimate is a_{ij}. Past research with the type of variables of interest may indicate what magnitude of salient loadings is to be expected. For example, in the area of personality-instrument development, evidence (Comrey & Montag, 1982; Oswald & Velicer, 1980; Velicer, DiClemente, & Corriveau, 1984; Velicer, Govia, Cherico, & Corriveau, 1985; Velicer & Stevenson, 1978) indicates that the magnitude of salient loadings is affected by the response format. A true–false response format will typically yield salient loadings in the low .40 range, whereas a Likert response format (five or seven choices) will typically yield higher (.60 range) component loadings.

This equation can also be used to evaluate the quality of a published analysis. The quality of the sample solution could be assessed by calculating Y based on the sample-size value used and the a_{ij} values obtained.

Discussion

The purpose of this study was to examine the conditions under which sample component patterns become stable with respect to their population patterns. We examined the effect of four factors (N, p, m, and a_{ij}). On the basis of the popularity of current sample-size rules, one would have expected (a) sample size and (b) number of variables to be of primary importance in determining comparability. Only sample size, however, was of major importance. In addition, component saturation (the magnitude of component loadings) was the factor that had the greatest impact. At the lowest saturation level (a_{ij} = .40), sample size was most clearly important in determining comparability. At higher saturation levels (.60 and .80), once a certain minimum sample size was achieved, further improvements were small.

The results obtained in this study provide little support for current rules. The most popular rules involve an N-to-p ratio and were clearly not substantiated. The rules differ in the recommended ratio of observations to variables; however, all rules require more observations as the number of variables increases. Results from this empirical investigation imply the opposite relation. Larger variable sets always possessed the smaller difference (g^2 value) between sample and population patterns at any sample-size level. The cutoff value of g^2 = .01 was generally attained at lower sample sizes for larger variable sets. The concept that more observations are needed as the number of variables increases is clearly incorrect.

One rule (Cattell, 1978) suggests that sample size be determined as a function of the number of expected factors. Increasing the number of components for a given number of variables

Table 4
Average g^2, Type I Error, Type II Error, and Kappa for 80 Patterns With .80 Loadings

Pattern	p	m	N	g^2	Type I error	Type II error	Kappa
1	36	3	50	.0093	0	0	1.00
2	36	3	100	.0039	0	0	1.00
3	36	3	150	.0028	0	0	1.00
4	36	3	200	.0021	0	0	1.00
5	36	3	300	.0015	0	0	1.00
6	36	3	500	.0007	0	0	1.00
7	36	3	1,000	.0003	0	0	1.00
8	72	3	100	.0044	0	0	1.00
9	72	3	150	.0024	0	0	1.00
10	72	3	200	.0022	0	0	1.00
11	72	3	300	.0014	0	0	1.00
12	72	3	500	.0008	0	0	1.00
13	72	3	1,000	.0004	0	0	1.00
14	36	6	50	.0109	0	0	1.00
15	36	6	100	.0047	0	0	1.00
16	36	6	150	.0029	0	0	1.00
17	36	6	200	.0025	0	0	1.00
18	36	6	300	.0015	0	0	1.00
19	36	6	500	.0009	0	0	1.00
20	36	6	1,000	.0005	0	0	1.00
21	72	6	100	.0048	0	0	1.00
22	72	6	150	.0031	0	0	1.00
23	72	6	200	.0023	0	0	1.00
24	72	6	300	.0016	0	0	1.00
25	72	6	500	.0009	0	0	1.00
26	72	6	1,000	.0005	0	0	1.00
27	108	6	150	.0027	0	0	1.00
28	108	6	200	.0021	0	0	1.00
29	108	6	300	.0016	0	0	1.00
30	108	6	500	.0009	0	0	1.00
31	108	6	1,000	.0004	0	0	1.00
32	144	6	150	.0031	0	0	1.00
33	144	6	200	.0023	0	0	1.00
34	144	6	300	.0015	0	0	1.00
35	144	6	500	.0009	0	0	1.00
36	144	6	1,000	.0005	0	0	1.00
37	36	9	50	.0099	0	0	1.00
38	36	9	100	.0045	0	0	1.00
39	36	9	150	.0033	0	0	1.00
40	36	9	200	.0022	0	0	1.00
41	36	9	300	.0016	0	0	1.00
42	36	9	500	.0010	0	0	1.00
43	36	9	1,000	.0005	0	0	1.00
44	72	9	100	.0047	0	0	1.00
45	72	9	150	.0031	0	0	1.00
46	72	9	200	.0024	0	0	1.00
47	72	9	300	.0016	0	0	1.00
48	72	9	500	.0010	0	0	1.00
49	72	9	1,000	.0005	0	0	1.00
50	108	9	150	.0031	0	0	1.00
51	108	9	200	.0022	0	0	1.00
52	108	9	300	.0016	0	0	1.00
53	108	9	500	.0009	0	0	1.00
54	108	9	1,000	.0005	0	0	1.00
55	144	9	150	.0034	0	0	1.00
56	144	9	200	.0025	0	0	1.00
57	144	9	300	.0016	0	0	1.00
58	144	9	500	.0009	0	0	1.00
59	144	9	1,000	.0005	0	0	1.00

Table 4 (*continued*)

Pattern	p	m	N	g^2	Type I error	Type II error	Kappa
60	108	18	150	.0034	0	0	1.00
61	108	18	200	.0026	0	0	1.00
62	108	18	300	.0017	0	0	1.00
63	108	18	500	.0010	0	0	1.00
64	108	18	1,000	.0005	0	0	1.00
65	144	18	150	.0035	0	0	1.00
66	144	18	200	.0026	0	0	1.00
67	144	18	300	.0017	0	0	1.00
68	144	18	500	.0010	0	0	1.00
69	144	18	1,000	.0005	0	0	1.00

affected comparability at the .40 component saturation level, but it seems to be the result of a smaller p/m ratio. If the ratio (p/m) remains constant, along with saturation and sample size, the value of g^2 remains almost unchanged as m increases. The relation between m and the N required to attain acceptable comparability was not exponential, as Cattell's (1978) rule suggests. Under the least well-defined (low p/m ratio and low component saturation) conditions, a sample size of 300 to 450 would be required for one to observe acceptable comparability between patterns.

Researchers recommending rules that suggest obtaining a maximum, minimum, or specific number of observations may find some support for their suggestions in these results. The results obtained, however, show that the sample size required to reach an acceptable level of comparability between patterns varies under the experimental conditions used. That is, the specification of one sample-size level as a universal value will overestimate the number of observations required to obtain acceptable comparability under some conditions and underestimate the number of observations under other conditions.

An examination of the results associated with each of the four comparison statistics (g^2, kappa, Type I error, and Type II error) used in this study reveals a consistent pattern. If a pattern was well-defined with respect to component saturation (a_{ij} = .60

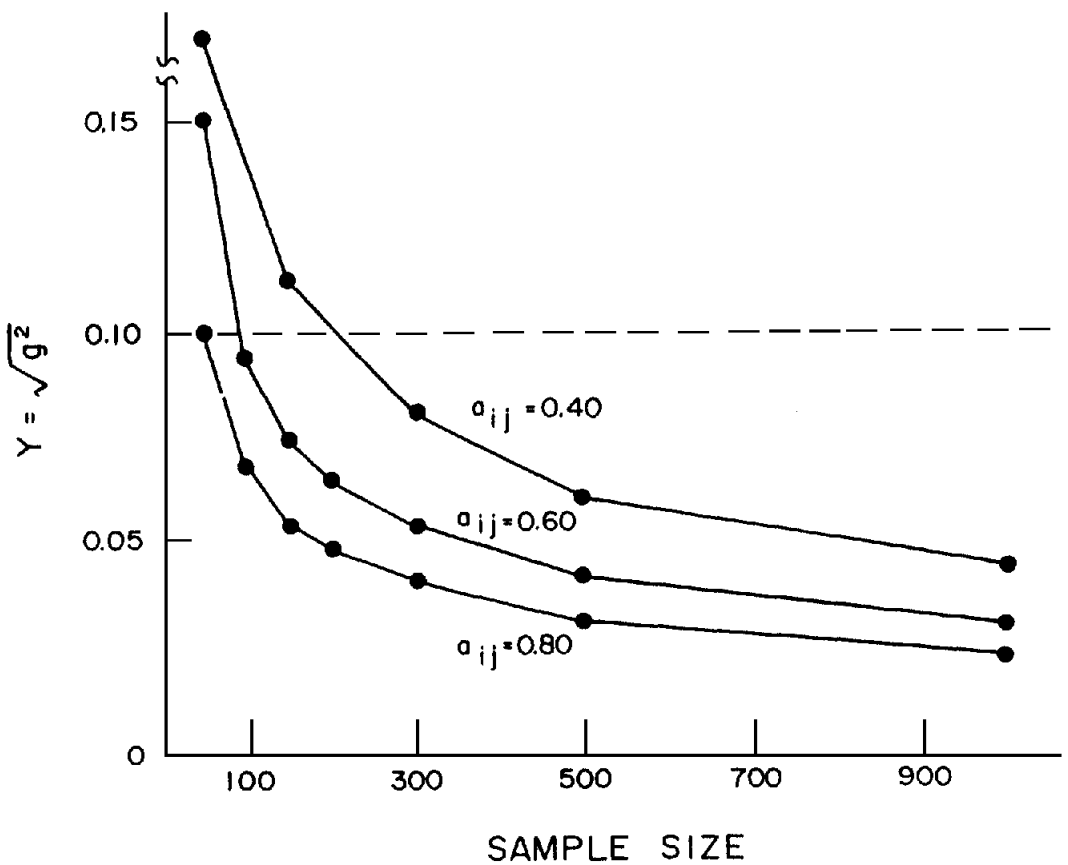

Figure 1. Relation between the square root of g^2 and sample size for three levels of saturation.

and $a_{ij} = .80$), then the number of variables, the number of components, and the sample size were not strongly related to the comparability of population and sample component patterns. When the saturation of a component was low ($a_{ij} = .40$), sample size had an effect on comparability. A second important factor was the number of variables defining the component (p/m). The better a component pattern was defined (high p/m ratio), the more accurately it reproduced the population component pattern. A low p/m ratio required a larger sample size to reproduce the population pattern.

The pattern of Type I and Type II errors occurring at the low saturation level ($a_{ij} = .40$) revealed a tendency for a component to become underdefined rather than overdefined. That is, there were more observed variables not identified as measuring a component (Type II error) than there were observed variables incorrectly identified as measuring a component (Type I error).

Further evidence supporting the importance of saturation level is provided by examining the conditions under which successful matching between the sample and population varimax-rotated patterns was a problem (see Table 2). Problems occurred only at the .40 saturation level. At this a_{ij} level, the effect of sample size and the number of variables per component (p/m) was also evident. Generally, problems in matching occurred when (a) sample sizes were small ($N = 50$ to 300), (b) there were few variables per component ($p/m = 4$ to 8), or (c) both.

Although these results were derived by using principal-components analysis, evidence suggests that a similar pattern would have resulted if a factor analytic procedure had been used. Velicer et al. (1982), in addition to finding that the solutions obtained from the two procedures differ minimally, also reported that the match of a sample factor pattern to its population pattern was quite good at the $N = 144$ level. Boomsma (1982), working with a structural equation model, suggested that such analyses should not be performed with fewer than 100 observations and that a sample size of 200 should provide more than adequate results. The sample-size values suggested by Velicer et al. (1982) and Boomsma (1982) fall within the range recommended here.

In summary, component saturation was the major factor in determining comparability between sample and population component patterns. At the lowest component saturation level used (.40), the effects of sample size and the number of variables per component became most evident. A good match to the population pattern was attained across all conditions when the sample component pattern was well-defined ($a_{ij} = .80$). Sample component patterns possessing moderate component saturation (.60) provided a good fit to the population pattern across conditions when sample size was greater than or equal to 150 observations. Weakly defined components ($a_{ij} = .40$, low p/m ratio) provided a good match only when sample size was in the range of 300 to 400 observations. Although these results are in direct contradiction to a wide class of rules of thumb in this area, they are consistent with a viewpoint based on the stability of the correlation coefficient.

Recommendations to the Applied Researcher

The applied researcher can use the results obtained here to maximize the chances of obtaining and interpreting a solution that best represents its population pattern. Given the importance of component saturation in determining comparability, the researcher, prior to an analysis, should select variables that will be good markers for a component—that is, variables that clearly should define a particular component and will load highly. If an a priori estimate of saturation level is difficult, many variables (10 or more) thought to represent a particular construct (component) should be selected. If these conditions can be accurately stipulated by the researcher beforehand, a sample size of 150 observations should be sufficient to obtain an accurate solution.

Following an analysis, the component pattern itself can be assessed with respect to the number of variables defining a component and with respect to the magnitude of component loadings. If components possess four or more variables with loadings above .60, the pattern may be interpreted whatever the sample size used. Similarly, a pattern composed of many variables per component (10 to 12) but low loadings ($a_{ij} = .40$) should be an accurate solution at all but the lowest sample sizes ($N < 150$). If a solution possesses components with only a few variables per component and low component loadings, the pattern should not be interpreted unless a sample size of 300 or more observations has been used. Replication is strongly suggested if these conditions occur when the sample size is fewer than 300 observations. Further, the prediction equation provided (Equation 3) may be used to calculate the approximate numeric fit of the sample pattern to the population pattern. This equation can be used prior to, following, or prior to and following an analysis.

References

Aleamoni, L. M. (1973). Effects of sample size on eigenvalues, observed communalities, and factor loadings. *Journal of Applied Psychology, 58,* 266–269.

Arrindell, W. A., & van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement, 9,* 165–178.

Baggaley, A. R. (1982). Deciding on the ratio of the number of subjects to number of variables in factor analysis. *Multivariate Experimental Clinical Research, 6,* 81–85.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality Study and Group Behavior, 1,* 23–33.

Bobko, P., & Schemmer, F. M. (1984). Eigen value shrinkage in principal component based factor analysis. *Applied Psychological Measurement, 8,* 439–451.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Joreskorg & H. Wold (Eds.), *Systems under indirect observation* (Part 1, pp. 149–173). Amsterdam: North-Holland.

Brislin, R. W., Lonner, W. J., & Thorndike, R. M. (1974). *Cross-cultural research methods.* New York: Wiley.

Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist.* New York: Harper & Row.

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum Press.

Comrey, A. L. (1973). *A first course in factor analysis.* New York: Academic Press.

Comrey, A. L. (1978). Common methodological problems in factor analytic studies. *Journal of Consulting and Clinical Psychology, 46,* 648–659.

Comrey, A. L., & Montag, I. (1982). Comparison of factor analytic results with two-choice and seven-choice personality item formats. *Applied Psychological Measurement, 6,* 285–289.

Driel, O. P.Van. (1978). On various causes of improper solutions in maximum likelihood factor analysis. *Psychometrika, 43,* 225–243.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: Wiley.

Glass, G., & Taylor, P. A. (1966). Factor analytic methodology. *Review of Educational Research, 36,* 566–587.

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

Guertin, W. H., & Bailey, J. P. (1970). *Introduction to modern factor analysis.* Ann Arbor: Edwards Brothers.

Guilford, J. P. (1954). *Psychometric methods.* New York: McGraw-Hill.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Grablowsky, B. J. (1979). *Multivariate data analysis.* Tulsa, OK: Petroleum.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24,* 417–441, 498–520.

Humphreys, L. G., Ilgen, D., McGrath, D., & Montanelli, R. (1969). Capitalization on chance in rotation of factors. *Educational and Psychological Measurement, 29,* 259–271.

Jackson, D. N., & Chan, D. W. (1980). Maximum-likelihood estimation in common factor analysis: A cautionary note. *Psychological Bulletin, 88,* 502–508.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika, 23,* 187–200.

Kunce, J. T., Cook, W. D., & Miller, D. E. (1975). Random variables and correlational overkill. *Educational and Psychological Measurement, 35,* 529–534.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement of categorical data. *Biometrics, 33,* 159–174.

Light, R. J. (1971). Measures of response agreement for qualitative data: Some generalizations and alternatives. *Psychological Bulletin, 76,* 365–377.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott, Foresman.

Loo, R. (1983). Caveat on sample sizes in factor analysis. *Perceptual and Motor Skills, 56,* 371–374.

Marascuilo, L. A., & Levin, J. R. (1983). *Multivariate statistics in the social sciences.* Monterey, CA: Brooks/Cole.

Montanelli, R. G., Jr. (1975). A computer program to generate sample correlation and covariance matrices. *Educational and Psychological Measurement, 35,* 195–197.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Oswald, W. T., & Velicer, W. F. (1980). Item format and the structure of the Eysenck Personality Inventory: A replication. *Journal of Personality Assessment, 44,* 283–288.

Press, S. J. (1972). *Applied multivariate analysis.* New York: Holt, Rinehart & Winston.

Pruzek, R. M., & Rabinowitz, S. N. (1981). A simple method for exploratory structural analysis. *American Educational Research Journal, 18,* 173–189.

Rummel, R. J. (1970). *Applied factor analysis.* Evanston, IL: Northwestern University Press.

Steiger, J. H., & Schönemann, P. H. (1978). A history of factor indeterminacy. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 136–178). San Francisco: Jossey-Bass.

Tucker, L. R., Koopman, R. F., & Linn, R. L. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika, 34,* 421–459.

Velicer, W. F. (1974). A comparison of the stability of factor analysis, principal component analysis, and rescaled image analysis. *Educational and Psychological Measurement, 34,* 563–572.

Velicer, W. F. (1976). The relation between factor score estimates, image scores, and principal component scores. *Educational and Psychological Measurement, 36,* 149–159.

Velicer, W. F. (1977). An empirical comparison of the similarity of principal component, image, and factor patterns. *Multivariate Behavioral Research, 12,* 3–22.

Velicer, W. F., DiClemente, C. C., & Corriveau, D. P. (1984). Item format and the structure of the Personal Orientation Inventory. *Applied Psychological Measurement, 8,* 409–419.

Velicer, W. F., & Fava, J. L. (1987). An evaluation of the effects of variable sampling on component, image, and factor analysis. *Multivariate Behavioral Research, 22,* 193–209.

Velicer, W. F., Govia, M. J., Cherico, N. P., & Corriveau, D. P. (1985). Item format and the structure of the Buss-Durkee Hostility Inventory. *Aggressive Behavior, 11,* 65–82.

Velicer, W. F., & Jackson, D. N. (1987). *Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure.* Manuscript submitted for publication.

Velicer, W. F., Peacock, A. C., & Jackson, D. N. (1982). A comparison of component and factor patterns: A Monte Carlo approach. *Multivariate Behavioral Research, 17,* 371–388.

Velicer, W. F., & Stevenson, J. F. (1978). The relation between item format and the structure of the Eysenck Personality Inventory. *Applied Psychological Measurement, 2,* 293–304.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research, 17,* 253–269.

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

Received August 25, 1986
Accepted August 17, 1987 ■