Open access • Journal Article • DOI:10.1177/014662169201600202

Replication as a Rule for Determining the Number of Clusters in Hierarchial Cluster Analysis — Source link

John E. Overall, Kevin N. Magee

Institutions: University of Texas at Austin

Published on: 01 Jun 1992 - Applied Psychological Measurement (SAGE Publications)

Topics: Cluster (physics)

Replication as a Rule for Determining the Number of Clusters in Hierarchial Cluster Analysis

John E. Overall and Kevin N. Magee
University of Texas Medical School

A single higher-order cluster analysis can be used to group cluster mean profiles derived from several preliminary analyses. Replication is confirmed when each higher-order cluster contains one cluster mean profile from each of the several preliminary analyses. This study evaluated the utility of replication as a stopping rule in hierarchical cluster analysis. Replication defined by higher-order clustering identifies the correct number of underlying populations that have distinct density regions in the multivariate measurement space. When increased within-population variance obliterates population distinctions, the replication criterion provides an underestimation of the actual number of latent populations. In the case of no true cluster structure or in the case of only two latent populations, chance replication can occur. Thus, replication suggested by higher-order cluster analysis is not a conservative test for the absence of a cluster structure, but it does provide valid evidence concerning the number of latent populations when several are present. *Index terms: cluster analysis, cluster means, hierarchical clustering, replication in cluster analysis, stopping rule in cluster analysis, validity of cluster analysis.*

In the behavioral sciences, the aim of cluster analysis is often to infer the nature of distinct underlying populations within a heterogeneous domain. "Population recovery" is the term used for the ability of clustering methods to segregate sample observations according to true differences in the underlying population memberships. Good population recovery requires determining the correct number of clusters, as well as establishing concordance between cluster and population memberships for the given number of clusters. However, much of the empirical work evaluating population recovery capabilities of cluster analysis procedures has assumed that the correct number of clusters is known (e.g., Blashfield, 1976; Milligan, 1980, 1981). The most serious unresolved problem in cluster analysis methodology concerns the questionable adequacy of criteria for determining the correct number of clusters (Milligan & Cooper, 1985).

Hierarchical cluster analysis (HCA) methods have contributed to the problem by routinely providing cluster solutions at all different hierarchical levels without adequate criteria for selecting among the alternative solutions. The hierarchical agglomerative procedures contained in the Statistical Package for the Social Sciences (SPSS; Norušis, 1986), Statistical Analysis System (SAS; SAS Institute, 1987), and BMDP biomedical computer library (Dixon, 1988) are the most widely available. These cluster analysis programs provide "fusion coefficients" that represent the increasing distances between units that are combined at successive hierarchical levels. SAS also has incorporated additional quantitative criteria for consideration as stopping rules in HCA.

Visual inspection of the pattern of fusion coefficients across hierarchical levels is similar to the use of the "scree criterion" for determining the number of factors to rotate in factor analysis. Mojena (1975) and Mojena and Wishart (1980) have attempted to define quantitative criteria for identifying a "significant jump" in the pattern of fusion coefficients, and some of these have been incorporated into specialized computer programs that unfortunately are not widely available (Wishart, 1982). Other authors have proposed alternative quantitative indices that tend to

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 16, No. 2, June 1992, pp. 119-128
© Copyright 1992 Applied Psychological Measurement Inc.
0146-6216/92/020119-10$1.75

approach a minimum or maximum value at the correct number of clusters (Davis & Bouldin, 1979) or that evaluate the fit between clustering and a priori structure (Hubert & Schultz, 1976). Textbook examples have been provided in which clear "shoulders" in the pattern of fusion coefficients coincide with known population differences (e.g., Aldenderfer & Blashfield, 1984; Jain & Dubes, 1988); however, in practice such clear demarcations are rare in the cases where cluster analysis seems needed most. Users are left with subjective choices among clustering levels, and a personal view of "meaningfulness" often dictates the solution that is reported.

Milligan and Cooper (1985) have provided comparative evaluations of approximately 30 internal criteria for stopping a cluster analysis, but simple higher-order cluster replication was not considered. Yet replication is generally considered a superior basis for scientific inference in most endeavors, and the use of replication as the criterion for selection of a final cluster solution is not a new approach. Lorr (1966) and Lorr, Klett, and McNair (1963) analyzed several subsets of data and retained as reliable clusters only those that replicated across different samples. Overall (1974; Overall, Hollister, Johnson, & Pennington, 1966; Overall & Rhoades, 1982) has pursued replication of cluster classification of psychiatric symptom profiles. The present paper discusses a procedure for using replication as a criterion and presents analyses based on artificial mixture data having known latent population composition.

Replication as a Stopping Rule in HCA

Cluster analysis can be used to group together multivariate profiles that have substantially similar form. Profiles within each cluster can be represented by a single mean profile that conveys the features distinguishing members of the cluster from members of other clusters. The term *higher-order cluster analysis* will be used to indicate cluster analysis applied to mean profiles for clusters derived from several independent preliminary analyses. A higher-order cluster analysis is examined as a way of determining whether the cluster mean profiles from the several preliminary analyses are similar enough to be indicative of replication. This is, of course, not the only way that replication of cluster analysis results can be examined (e.g., Breckenridge, 1989; McIntyre & Blashfield, 1980), but it is the method considered here.

Replication is suggested when cluster mean profiles from different preliminary analyses are similar enough to be grouped together by a higher-order analysis. *Perfect replication* results when higher-order clusters each contain exactly one cluster mean profile from each of the several preliminary analyses. For example, at the five-cluster level of K preliminary analyses, perfect replication would be realized if a higher-order cluster analysis produced five clusters each containing a single mean profile from each of the K preliminary analyses. The objective criterion of perfect replication is proposed to define the appropriate number of clusters (stopping rule) in HCA.

To use perfect replication as a stopping rule, first randomly split the total available sample of multivariate measurement profiles into K subsamples. Based on preliminary work with population recovery as a criterion, the total sample in this study was divided into four independent subsamples, although the strictness of the perfect replication criterion obviously increases with the number of preliminary analyses that contribute cluster mean profiles to the higher-order analysis. The aim here was to illustrate the method, rather than to prematurely restrict its generality.

A preliminary HCA is conducted on each of the independent subsamples to define solutions at hierarchical levels 2 through M, where M is a number substantially exceeding the potential number of replicating clusters. Next, the issue of perfect replication is examined. Begin at the two-cluster level and work progressively up the hierarchy, performing a single higher-order analysis of cluster mean profiles for each level separately. Replication of the two-cluster solution is evaluated at the two-cluster level of the higher-order analysis; replication at the three-cluster level is evaluated at the three-cluster level of the higher-order analysis; and so forth. Thus, a separate higher-order analysis must be accomplished on the cluster means derived from each level of the hierarchy in the several preliminary analyses. The appropriate hierarchical level to infer true underlying population differences is the highest level at which perfect replication can be documented.

The purpose is not to tie the proposed replication criterion to a particular form of cluster analysis nor to specify four subsamples as a required number. The intended purpose here was to demonstrate the utility of clustering of cluster means as an objective criterion for replication. The proposed algorithm can be summarized as follows:

1. Randomly partition a dataset consisting of N multivariate measurement vectors into K independent subsamples.
2. Hierarchically cluster each of the K subsample datasets and calculate cluster mean profiles separately at hierarchical levels $m = 1, 2, \ldots, M$.
3. At each hierarchical level m, subject the mK cluster mean profiles to a higher-order HCA and examine the replication at level m of that analysis, disregarding results at the other levels.
4. If none of the higher-order clusters at level m contains more than one mean profile from each of the K original analyses, conclude that perfect replication has been achieved at cluster level m.
5. Identify the highest value of m at which perfect replication is achieved and infer that the original dataset contains that number of latent populations.

In any content domain, some latent populations are likely to be more distinct and some more overlapping in their distributions on the measurements available for analysis. Obviously, in the presence of highly overlapping latent populations, it is possible that perfect replication will not be observed at any level. In other cases, some underlying populations will have distinct modes that can be recognized by the cluster analysis, whereas other populations cannot be discriminated. Thus, perfect replication is proposed as a criterion for defining a lower bound for the number of latent populations present within a heterogeneous domain.

Evaluation of the Replication Criterion

Any empirical multivariate sampling study must necessarily focus on only a subset of the many parameters that are of potential interest. In this study, the number of underlying populations from which the mixture samples were randomly drawn and the degrees of overlap among those populations were systematically varied. The population mean profiles were randomly varied from one dataset to the next to enhance generality. Parameters that were not varied included a fixed number of elements in the multivariate measurement profiles, the equal base rates for the latent populations, and the normality and statistical independence of the multiple measurements within the latent populations. Ward's (1963) method using squared Euclidean distance measures was employed because preliminary work by the authors provided evidence of superior population recovery by that method across different distance measures and across different methods of cluster analysis. Equal-sized populations and multivariate normality have been reported by others to favor the use of Ward's method (Aldenderfer & Blashfield, 1984). Again, the purpose was not to restrict utility of the cluster replication criterion to a single clustering method. The selection of a clustering method should be based on the type of data being analyzed and expectations concerning cluster structure.

Method

Data were generated to represent random samples from a mixture of two, three, four, or six multivariate normal populations with seven different levels of overlap in their sampling distributions. With 100 sampling replications for each combination of number of latent populations and degree of overlap, a total of 2,800 mixture datasets were generated for analysis. Because some latent populations can be expected to differ more than others in terms of the particular measurements available for analysis, a normal random number generator was employed to define the population means, as well as the sample observations, within each dataset separately. A 10-element mean profile was generated randomly to represent each latent population, and the sample data were generated by adding independent random normal deviates of specified variance to the 10 elements of the population mean profiles. The random normal deviate generator of the IMSL (IMSL, 1982) software library was used to generate both the population mean configuration and the data sampled from those populations for each of the 2,800 mixture datasets. The generated data represented 100 independent random samples from two, three, four, or six overlapping multivariate normal distributions with diagonal covariance structures. Differences among the population mean profiles introduced correlations among the measurements across the total sample, but it was assumed that measurements were selected to represent separate traits and that independent error alone separated individuals within truly homogeneous populations.

The degree of overlap in the sampling distributions for the underlying populations is critical for determining the "clusterability" of data. The variance of the independent random deviates, which were added to the population mean profiles to generate sample data, was controlled to produce different magnitudes of overlap for different series of analyses. Specifically, for each dataset, the within-population variance was scaled to be a constant fraction of the variance of the randomly generated population means on each of the 10 variables, and an intraclass correlation coefficient (Winer, 1971, p. 248) was calculated to describe the separation (or overlap) of the population distributions. Because of this method of scaling the within-population variances relative to the variability of the randomly generated population means, a constant average separation among the latent populations was assured on each variable. However, because the population means were randomly generated, it was quite possible for individual pairs of populations to be indistinguishable within datasets consisting of samples from more than two underlying populations.

As the within-population variance increased, the overlap in sampling distributions eventually obliterated any detectable differences among the latent populations. One additional monte carlo series of 100 datasets was run to evaluate the performance of the perfect replication criterion when samples were drawn from a single homogeneous population without cluster structure. Those results are considered the limiting case of increasing variance for the two, three, four, and six population conditions. They are reported as representing an intraclass correlation of 0.

Monte carlo methods were used to evaluate the utility of higher-order clustering for defining a lower bound on the number of latent populations that are present in a heterogeneous domain. For each of the combinations of number of randomly generated population means and specified degrees of overlap, a series of 100 mixture datasets was analyzed. Each dataset involved independently generated population mean profiles to which the random sampling error was added in specified magnitude. Each dataset consisted of a total of 192 10-element sample profiles that were then randomly divided into four subsamples of size $n = 48$. Four preliminary cluster analyses were run on the four independent subsamples using the hierarchical agglomerative cluster analysis program of SPSS (Norušis, 1986), based on Ward's (1963) method with squared Euclidean distance measures. Cluster mean profiles were calculated for hierarchical levels 2 through 8 in the four preliminary analyses, and a higher-order cluster analysis was performed on the mean profiles at each hierarchical level to identify the highest level of perfect replicability.

Results

The frequencies with which the perfect replication criterion selected different hierarchical levels in the presence of different numbers of underlying latent populations and different degrees of (average) overlap in their sampling distributions are presented in Table 1. The intraclass correlation coefficients, which were calculated from sample data, represent the ratio of true population differences to true variance plus sampling variance and were calculated separately for the 10 elements of the individual data profiles and then averaged. When mixture samples involved three, four, or six latent populations with reasonably discriminable distributions, as indicated by the intraclass correlation coefficients greater than .7, the replication criterion was essentially perfect as a stopping rule. As the degree of overlap among the randomly generated

Table 1
Replication Frequencies at Different Cluster Levels and Different Levels of Intraclass *R*

Populations and	Number of Clusters							
Intraclass *R*	0	2	3	4	5	6	7	8
Six Latent Populations								
.930	0	0	0	0	0	100	0	0
.775	0	0	0	1	0	99	0	0
.660	0	3	1	7	20	69	0	0
.553	0	11	20	21	16	31	1	0
.461	5	27	23	27	13	5	0	0
.232	23	62	14	1	0	0	0	0
.114	38	48	13	1	0	0	0	0
.000	76	22	2	0	0	0	0	0
Four Latent Populations								
.895	0	0	0	98	2	0	0	0
.790	0	0	1	99	0	0	0	0
.677	0	0	3	97	0	0	0	0
.571	0	0	8	92	0	0	0	0
.478	0	6	26	66	2	0	0	0
.242	14	53	28	5	0	0	0	0
.119	47	43	7	3	0	0	0	0
.000	76	22	2	0	0	0	0	0
Three Latent Populations								
.904	0	0	100	0	0	0	0	0
.815	0	0	100	0	0	0	0	0
.698	0	0	99	1	0	0	0	0
.560	0	2	97	1	0	0	0	0
.531	0	2	94	4	0	0	0	0
.272	9	43	47	1	0	0	0	0
.138	35	48	17	0	0	0	0	0
.000	76	22	2	0	0	0	0	0
Two Latent Populations								
.930	0	93	5	2	0	0	0	0
.856	0	90	7	3	0	0	0	0
.772	0	98	2	0	0	0	0	0
.686	0	94	5	1	0	0	0	0
.605	0	95	2	3	0	0	0	0
.363	0	86	12	2	0	0	0	0
.209	15	76	8	1	0	0	0	0
.000	76	22	2	0	0	0	0	0

latent populations increased, the frequency with which some populations could not be separated by cluster analysis increased. Thus, in the presence of highly overlapping populations, the replication criterion tended to underestimate the actual number of latent populations. When there were only one or two populations, replicability was observed at levels beyond the actual number of latent populations. This is explained by the fact that chance agreement with any particular cluster pattern has a higher probability when the number of profiles to be sorted is small. Nevertheless, even with a small number of latent populations, these results show that the higher-order replication criterion generally performed well.

The final row of each section of Table 1 reproduces results obtained in an application of the higher-order replication criterion to four samples from a single homogeneous population, which is equivalent to a theoretical intraclass coefficient of 0.0. In this case of no true underlying population differences, the four-sample replication criterion incorrectly suggested the presence of more than a single population in 24% of the homogeneous datasets. Thus, higher-order replication across four subsamples appears inadequate as a test for the absence of any true population differences. However, if the perfect replication criterion is met at hierarchical level 3 or above, there is basis for confidence that the data derive from a mixture of underlying populations, even in the presence of substantial population overlap. This is evident in the fact that only 2 of 100 datasets drawn randomly from a single population resulted in perfect replication at the three-cluster level and none at a level beyond that.

Table 2
Accuracy of Classification of the Number of Underlying Latent Populations for Various Levels of Intraclass Correlation (*R*)

R	% Under	% Correct	% Over
.90–1.00	0.0	97.7	2.3
.80–.89	0.0	96.0	4.0
.70–.79	.7	98.7	.6
.60–.69	6.8	90.8	2.4
.50–.59	20.0	78.5	1.5
.40–.49	63.5	35.5	1.0
.10–.39	68.0	29.0	3.0
.00–.09		76.0	24.0

Table 2 summarizes the probabilities of identifying correctly the actual number of latent populations (two, three, four, or six) as a function of population overlap. At no level of overlap among two or more underlying populations, as represented by intraclass correlation coefficients above 0, did the probability of overestimation exceed .05. For intraclass coefficients greater than .7, higher-order replication based on four independent subsamples had a hit rate exceeding 95% for correct identification of the actual number of latent populations represented in the mixture datasets. For intraclass coefficients below .7, the frequencies of underestimation progressively increased as the degree of population overlap increased.

Discussion and Conclusions

The adequacy of higher-order cluster analysis as a criterion for inferring the number of underlying latent populations is demonstrated for the particular conditions that were examined here—which included sampling from equal size multivariate normal populations having diagonal covariance structure. These conditions were selected to represent what might be considered the ideal design for cluster analysis research. Normality of error distributions is a generally accepted condition of measurements within naturally-occurring homogeneous populations, although appropriate scaling of measurements may be required to reflect this. Based on this expectation, univariate "mixture model" methods consider departures from normality to be evidence of heterogeneity.

Because heterogeneous correlations among profile elements imply differential weightings of primary underlying dimensions of individual differences in profile similarity indices, cluster analysis research generally uses measurements that are both conceptually and statistically independent. Although complete within-population independence of profile elements may not be achieved in practice, it is reasonable to take into account when simulating data for evaluating methods. Although measurements are uncorrelated within underlying populations, differences in the population mean profiles will produce correlations across samples from a mixture of the populations. Assuming that conceptually distinct measurements should be uncorrelated within homogeneous populations, independence of measurements within clusters can be viewed as justification for homogeneity, much as univariate mixture analysis uses normality of score distributions as a criterion for homogeneity. Thus, independence of multiple profile elements within underlying populations is a reasonable model to assume in generating sampling data.

The assumption of equal population base rates is less easily justified. Although there are theoretical grounds to justify multivariate normality and diagonal covariance structure within latent populations, there is indeed no theoretical basis for assuming that underlying populations should be of equal size, as was the case in this investigation. Equal population base rates were used here because the interest was in examining the higher-order cluster replication criterion, not in comparing different clustering methods. Ward's (1963) method was selected because it has been reported to be especially prone to define equal-size clusters. As will be noted, it should probably be preferred for the higher-order analysis, regardless of the rationale for selection of a clustering procedure at the preliminary data stage.

Two issues regarding equal-size clusters are involved here. One pertains to the number of clusters obtained in application to the original mixture samples, and the other pertains to replication confirmed by higher-order analysis of cluster mean profiles. If the relative sizes of the underlying latent populations are substantially different, then it is likely that the smaller populations would be represented adequately in sample data less often, and the perfect replication criterion would tend to underestimate the true number of latent populations. A cluster method other than Ward's (1963) might have some advantages in that case, but the general tendency to miss small-size populations still would be present. On the other hand, the perfect replication criterion demands equal-size clusters in the higher-order analysis of cluster means. As long as replication is judged by higher-order clusters that contain exactly one cluster mean from each preliminary analysis, a method of cluster analysis that favors equal-size clusters is logically required for the higher-order analysis, regardless of the method used for the several preliminary analyses. However, if the possibility of substantially unequal population base rates is admitted, and a clustering method that readily defines unequal-size clusters is used on independent samples at the preliminary stage, the definition of perfect replication might be modified to require only that no more than one cluster mean from a single preliminary analysis be included in each higher-order cluster, even though every preliminary analysis might not be represented in each higher-order cluster. In that case, the higher-order analysis would have to combine results from different hierarchical levels of the several preliminary analyses. A key to whether that might be considered important would be the appearance of clusters of highly variable sizes within all, or most, of the several preliminary analyses. Obviously, there is room for additional work in exploring the variations that are possible in real data and their implications for evaluating replicability of cluster results.

The results reported here, however, confirm that under reasonably representative conditions the replication demonstrated by clustering mean profiles from several preliminary cluster analyses of independent subsets of multivariate profiles is an adequate criterion for inferring the correct number of underlying populations when those latent populations are reasonably well separated. As the overlap in the sampling distributions increases, there is a tendency for the replication criterion to underestimate the number of latent populations. However, randomly generated population means may be so similar that increasing variance might render the overlapping distributions indistinguishable in some cases. A similar problem of underestimation has been reported for other stopping rules in the presence of "weakly clustered data" (Jain & Dubes, 1988, pp. 187–188). It is inappropriate to fault failure of cluster replication in cases where underlying populations are not discriminably different. True multimodality is necessary for cluster analysis to recognize consistently different underlying populations.

The present results show that higher-order cluster analysis is an inadequate basis for determining that no true population differences are present. The relative inadequacy of this method as a criterion when there are no more than two underlying populations reflects the same weakness. Given purely random data, HCA will produce clusters. When a higher-order analysis is limited to sorting only eight mean profiles at the two-cluster level, a random sort has substantial likelihood of assigning one and only one mean profile from each preliminary analysis to each of two higher-order clusters. The problem is compounded by the fact that cluster analysis at the two-cluster level tends to partition available samples into contrasting groups, even if no true cluster structure is present. The tendency to partition the sample distributions into contrasting clusters is enhanced by the presence of correlations among the profile elements. This enhances the likelihood that cluster means from different samples will have patterns similar enough to be sorted into contrasting groups by a forced partition at the second or third hierarchical level, even if no true population differences are present.

The weakness of overestimating the number of latent populations can be removed by increasing the number of preliminary analyses that contribute cluster means to the higher-order analyses, but doing that enhances the probability of failing to distinguish clusters that may not be separated well. Perfect replication is less likely to occur by chance across a larger number of preliminary analyses. To examine this phenomenon, datasets of N = 192 were randomly generated simulating sampling in equal numbers from a mixture of two overlapping 10-dimensional latent populations. The within-population variances were scaled to produce an intraclass correlation R = .686 (see Table 1). However, each total sample was randomly partitioned into six rather than four subsamples on which six preliminary hierarchical cluster analyses were accomplished. The perfect replication criterion correctly identified the two-cluster level in 99 out of 100 runs, with perfect replication at the three-cluster level in the remaining case. When this analysis was repeated using six preliminary subsample analyses with a greater population overlap (R = .209), the perfect replication criterion across the six preliminary cluster analyses correctly identified the two-cluster level in only 53 out of 100 runs. The strict replication criterion was not satisfied at any hierarchical level in 43 of the 100 analyses.

The practical issues are what number of preliminary analyses to recommend and how to accomplish multiple preliminary analyses with limited datasets. Fortunately, no more than four preliminary analyses appear required to avoid the problem of overestimating the number of distinct latent populations when that number exceeds two. Recognizing the danger of failing to identify population differences that are present by use of an overly strict replication criterion, four subsample analyses are recommended to evaluate perfect replication, thus accepting a modest chance of overestimating the number of latent populations when there are no more than two. Overestimation is not a significant problem when the number of underlying populations exceeds two. However, underestimation is likely when several latent populations are substantially overlapping in their distributions on the measurements available for analysis.

The conservatism of the replication criterion can be relaxed by decreasing the number of subsample analyses to three. This will produce better estimation for a larger number of overlapping latent populations, but it enhances the probability of chance replication of false clusters that do not recover the true population differences.

This is why the four subsample replication criterion is recommended, although it tends to be conservative when several latent populations with substantially overlapping distributions are present.

In cases in which the number of latent populations is small, the sample sizes for the preliminary analyses need not be as large as required to support a larger number of clusters. With expectation of reasonably comparable base rates for the different latent populations, minimum sample sizes for the preliminary analyses should be 8 to 10 times the number of latent populations. Thus, if empirical results, or a priori considerations, suggest only two or three distinguishable populations, preliminary cluster analyses on subsamples of size $n = 30$ should be adequate.

If the total available dataset is too small to permit partitioning into mutually exclusive subsamples, random sampling with replacement from a total $n = 100$ or more should be adequate to verify the presence of two to four latent populations. The strategy of minimizing duplication has been used, rather than strict random sampling. This can be done by randomly splitting the total available sample into two equal parts, then randomly splitting each of the two parts and combining those halves into two additional samples having no more than 50% overlap with the first two subsamples. In this way, two pairs of subsamples are mutually exclusive and the other two are minimally overlapping.

Although perfect replication was emphasized here as a stopping rule for HCA, something short of perfect replication (as defined here) should serve as a useful basis for inferring the population reality of preliminary clusters that do replicate across four independent subsamples. ''Near-perfect replication'' is present in higher-order clusters that contain cluster means from at least three of the four preliminary analyses and no more than two cluster means from any one of the preliminary analyses. In this case, one or more of the higher-order clusters may be discounted as failing to provide adequate evidence of replication. Obviously, perfect replication is preferred, but in circumstances in which several latent populations are reasonably expected to be highly overlapping in terms of the available measurements, the weaker replication criterion may be required. Because it involves greater probability of chance replication, at least four latent populations should be confirmed by the higher-order analysis when the weaker criterion is to be relied on as a basis for inference. That is, near-perfect replication is inadequate to support the reality of only two or three underlying latent populations, because that has an unacceptable likelihood of occurring by chance.

Replication is a trusted foundation for scientific inference, and higher-order clustering is a convenient, logical, and objective way to evaluate replication. Except in the case of a mixture of only two latent populations, four preliminary subsample analyses should provide an adequate basis for defining a lower bound on the number of distinguishable latent populations. It is a lower bound primarily because some latent populations may be indistinguishable in terms of the available measurements. That was particularly true when population mean profiles were randomly generated, as they were in this evaluation of the higher-order clustering approach to replication.

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Beverly Hills CA: Sage Publications.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin, 83,* 277–288.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24,* 147–162.

Davis, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI, 1,* 224–227.

Dixon, W. J. (1988). *BMDP statistical software manual.* Berkeley CA: University of California Press.

Hubert, L. J., & Schultz, J. (1976). Quadratic assignment as a general data-analysis strategy. *British Journal of Mathematical and Statistical Psychology, 29,* 190–241.

IMSL (1982). *IMSL library reference manual* (9th ed.). Houston TX: IMSL Inc.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for* *clustering data.* Englewood Cliffs NJ: Prentice Hall.

Lorr, M. (Ed.). (1966). *Explorations in typing psychotics.* New York: Pergamon.

Lorr, M., Klett, C. J., & McNair, D. M. (1963). *Syndromes of psychosis.* New York: Pergamon Press.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 2,* 225–238.

Milligan, G. W. (1980). An examination of the effects of six types of error perturbation on fifteen clustering algorithms. *Psychometrika, 45,* 325–342.

Milligan, G. W. (1981). A Monte-Carlo study of 30 internal criterion measures for cluster analysis. *Psychometrika, 46,* 197–195.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159–179.

Mojena, R. (1975). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20,* 359–363.

Mojena, R., & Wishart, D. (1980). Stopping rules for Ward's clustering method. *Proceedings of COMPSTAT 1980.* Wurzburg, Germany: Physika-Verlag.

Norušis, M. J. (1986). *SPSS/PC+ advanced statistics.* Chicago: SPSS Inc.

Overall, J. E. (1974). The Brief Psychiatric Rating Scale in psychopharmacology research. In P. Pichot and R. Oliver-Martin (Eds.), *Psychological measurements in psychopharmacology: Modern problems in pharmacopsychiatry* (pp. 67–78). Basel: Karger.

Overall, J. E., Hollister, L. E., Johnson, M., & Pennington, V. (1966). Nosology of depression and differential response to drugs. *Journal of the American Medical Association, 195,* 946–948.

Overall, J. E., & Rhoades, H. M. (1982). Refinement of phenomenological classification in clinical psychopharmacology research. *Psychopharmacology, 77,* 24–30.

SAS Institute Inc. (1987). *SAS/STAT guide for personal computers: Version 6 edition.* Cary NC: Author.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58,* 236–244.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

Wishart, D. (1982). *CLUSTAN User Manual* (3rd ed., Supplement) [Computer program manual]. Edinburgh: Program Library Unit, Edinburgh University.

Author's Address

Send requests for reprints or further information to John E. Overall, Department of Psychiatry and Behavioral Sciences, University of Texas Medical School, Houston TX 77225, U.S.A.