Multivariate Behavioral Research, 1979, 14, 57-74

HIERARCHICAL CLUSTER ANALYSIS AND THE INTERNAL STRUCTURE OF TESTS

WILLIAM REVELLE
Northwestern University

ABSTRACT

Hierarchical cluster analysis is shown to be an effective method for forming scales from sets of items. The number of scales to form from a particular item pool is found by testing the psychometric adequacy of each potential scale. Higher-order scales are formed when they are more adequate than their component sub-scales. It is suggested that a scale's adequacy should be assessed by a new measure of internal consistency reliability, coefficient beta, which is defined as the worst split-half reliability of the test.

Comparisons with other procedures show that hierarchical clustering algorithms using this psychometrically based decisions rule can be more useful for scale construction using large item pools than are conventional factor analytic techniques.

A common problem in the social sciences is to construct scales or composites of items to measure constructs of theoretical interest and practical importance. This process frequently involves administering a battery of items from which those that meet certain criteria are selected. These criteria might be rational, empirical, or factorial (Goldberg, 1972). A similar problem is to analyze the adequacy of scales that already have been formed and to decide whether the putative constructs are measured properly. Both of these problems have been discussed in numerous texts, (*e.g.*, Guilford, 1954; Nunnally, 1967; Wiggins, 1973) as well as in myriad articles. Proponents of various methods have argued for the importance of face validity, discriminant validity, construct validity, factorial homogeneity, and theoretical importance. This paper will continue the debate by suggesting a new (or at least revised) estimate of factorial homogeneity and will outline a procedure for constructing scales using this and similar estimates.

Consider the following example: A group of items has been administered to some subjects. Each item is assumed to have some variance that is common with at least several other items, some unique variance, and some remaining variance that reflects moment to moment fluctuations on the part of the subjects. The average inter-item correlation of all the items is low and the number of subjects does not greatly exceed the number of items (if at all). Rather than consider all possible response patterns to these

Requests for reprints should be sent to William Revelle, Department of Psychology, Northwestern University, Evanston, Illinois 60201

items, subsets of the entire item pool are to be grouped into scales or indices. To be theoretically meaningful, these scales are to be factorially homogeneous. To be practically useful, they are to be independent of each other. The problem thus is how to partition the entire set of items into internally consistent and independent subsets. This problem may be thought of in terms of three separate questions: 1) how many scales should be formed; 2) how to assign items to these scales; and 3) what is the quality of these resulting scales. These are essentially questions of what to measure, how to measure it, and how well it is measured.

The solution proposed for all three of these questions is to apply principles of hierarchical cluster analysis to the problem of scale construction. Cluster analysis is a loosely defined set of procedures associated with the partitioning of a set of objects into non-overlapping groups or clusters, (Everitt, 1974; Hartigan, 1975). Although normally used to group objects, occasionally cluster analysis is applied to the problem of grouping variables and as such is similar to procedures of group factor analysis. (Loevinger, Gleser, and Dubois, 1953; Tryon and Bailey, 1970; Hartigan, 1975).

Hierarchical cluster analysis procedures are well known and have been reviewed recently by Everitt (1974), Hartigan (1975) and Blashfield (1976). Many of the major varieties of hierarchical clustering procedures have been incorporated into a computer package (CLUSTRAN) by Wishart (1969). Few of these procedures, however, have been geared to the psychometric problem of identifying item composites that are both internally consistent and relatively independent. If they have, they have not used psychometrically relevant decision rules or measures. It is possible, though, to combine psychometric principles with clustering procedures. This combination results in a simple but useful approach to scale construction.

Before it is possible to describe such a combination, however, it is necessary to outline the basic procedures of hierarchical clustering:

1) find the inter-item similarity matrix.
2) find the most similar pair of variables from this matrix.
3) combine these two variables into a new (composite) variable.
4) calculate the similarity of this composite variable with the remaining variables.

5) repeat steps 2-4 considering both initial variables and composites of variables.
6) stop the procedure when there are no more variables to combine or when some criterion has been reached.

The chief differences between clustering algorithms are: 1) how to define the initial similarity matrix; 2) how to calculate the similarity of a composite variable (cluster) with other variables or clusters; and 3) when to stop clustering.

In each of these three areas there is a natural solution for the formation of item composites or tests. Makers and users of tests are interested in two properties of tests: their intercorrelations with other tests and estimates of the test reliability. Thus, a reasonable inter-cluster similarity measure is either the correlation or the covariance between two clusters. Similarly, a reasonable way to combine clusters is to define the composite cluster as the sum of the unit-weighted items within each subcluster. Finally, a reasonable time to stop combining clusters is when some estimate of the internal consistency of the composite cluster is less than that of the component clusters.

The first step of hierarchical cluster analysis is to find the correlation matrix. The second step is to find and combine those two most similar variables. The simplest definition of similarity is the raw correlation coefficient. One that takes into account the range of possible correlations for a variable is the unattenuated correlation coefficient (the raw correlation divided by the geometric mean of the reliabilities of the variables). An initial estimate of the reliability can be the highest correlation that variable has with any other variable. This corrected similarity measure has the effect of identifying and clustering reciprocal pairs of variables (McQuitty & Koch, 1975), *i.e.,* those variables which have their highest correlations with each other.

The third step of hierarchical clustering is to combine this pair of variables and to calculate the similarity of this composite variable with the remaining variables (deleting the members of the composite). The correlation of the unweighted composite $x_1 + x_2$ with variable x_3 is the sum of the unit-weighted zero-order covariances divided by the geometric mean of the composite variance and the variance of x_3, *i.e.*

$$r_{(1+2)3} = (\sigma_{13} + \sigma_{23}) \,/\, \sqrt{(\sigma_3^2)(\sigma_1^2 + \sigma_2^2 + 2\sigma_{12})}$$

The unattenuated correlation of the cluster with other variables may be estimated by using coefficient alpha of the cluster as an estimate of the cluster reliability. An alternative view of the unattenuated correlation between cluster (A and B) is as the ratio of the average between cluster covariance ($\overline{\sigma}_{ij'}$) to the geometric mean of the average within cluster covariance $\sqrt{\overline{\sigma}_{ij} \cdot \overline{\sigma}_{i'j'}}$.

$$\begin{aligned}\tilde{r}_{AB} &= r_{AB}/\sqrt{\alpha_A\alpha_B} = C_{AB}/\sqrt{\alpha_A V_A \alpha_B V_B} = nm\overline{\sigma}_{ij'}/\sqrt{n^2\overline{\sigma}_{ij} \cdot m^2\overline{\sigma}_{i'j'}} \\ &= \overline{\sigma}_{ij'}/\sqrt{\overline{\sigma}_{ij}\overline{\sigma}_{i'j'}}\end{aligned}$$

The fourth step in hierarchical clustering is to find the next most similar pair of variables and to repeat the second and third steps until either there are no more variables to combine, or until some criterion has been reached. One such stopping criterion that has been suggested by Loevinger, Gleser and DuBois (1953) for nonhierarchical clustering and by Kulik, Revelle and Kulik (Note 2) for hierarchical procedures is to combine variables until coefficient alpha fails to increase; that is, until coefficient alpha of the combined cluster is less than that in either or both of the sub clusters. (This will be referred to in the rest of the text as the alpha clustering rule.)

A difficulty with this criterion is that although alpha is an upper bound of the percentage of test variance that may be associated with a general factor and is a lower bound of the percentage of test variance associated with the sum of all common factors, it is sometimes a very poor estimate of the general factor saturation of a test (Cronbach, 1951). It is well-known that if a test is "lumpy", or has several large group factors, then alpha can be large even though the percentage of test variance associated with a general factor is low or nonexistent.

An alternative estimate of the general factor saturation is to consider the *worst split-half reliability* estimate. Call this worst split-half *coefficient beta.* In the case of split halves (A and B) of equal length, then beta is $4\sigma_{AB}/\sigma^2_{(A+B)}$ where σ_{AB} is minimized. Since $\sigma^2_{(A+B)} = \sigma_A^2 + \sigma_B^2 + 2\sigma_{AB}$ is fixed for a test, minimizing σ_{AB} is the same as maximizing $\sigma^2_A + \sigma^2_B$. Thus, coefficient beta can be found by partitioning the test into 2 sub-tests such that the between-test covariance is minimized or that the sum of the within-test variances is maximized. In the more general case of split halves of unequal lengths, beta is defined to be the average between-half covariance times the total number of test items squared divided by the total variance, *i.e.*,

$$\beta = (n + m)^2 \bar{\sigma}_{ij'} / \sigma^2_{(A+B)}$$

where the split halves are of length n and m and the average between-half covariance ($\bar{\sigma}_{ij'}$) is minimized. While alpha is sensitive to components of variance within subtests as well as between subtests, beta is sensitive only to components of variance between subtests. Furthermore, since alpha is the mean of all split halves and beta is the worst split-half, alpha will always be greater than or equal to beta.

To better understand the relationship between these two indices of internal consistency and how they relate to the problem of estimating the amount of test variance due to a general factor, it is useful to consider hypothetical tests made up of homogeneous subtests of length n. Let r stand for the average correlation between the two subtests and r' represent the average correlation within each of these two subtests. The only component of variance contributing to r is the general factor saturation of each item; the components of variance contributing to r' are the general and group (subtest) factor saturations. The values of coefficients alpha and beta as well as the average item loading on the general and group factors for such a test are shown in Table 1 as a function of different values of r and n. For the purpose of this illustration,

Table 1
Coefficients *Alpha* and *Beta* as a Function of Test Length and General Factor Saturation of Items. (The Correlation Within Subtests is Set to .25.)

Factor Saturation		Sub-test Length 5		10		20	
general	group	α	β	α	β	α	β
$\sqrt{.25}$	$\sqrt{.00}$	.77	.77	.87	.87	.93	.93
$\sqrt{.20}$	$\sqrt{.05}$	.74	.67	.85	.76	.92	.82
$\sqrt{.15}$	$\sqrt{.10}$	.71	.55	.83	.63	.91	.69
$\sqrt{.10}$	$\sqrt{.15}$	.67	.40	.80	.47	.89	.52
$\sqrt{.05}$	$\sqrt{.20}$	.62	.22	.77	.27	.87	.30
$\sqrt{.00}$	$\sqrt{.25}$	.56	.00	.73	.00	.85	.00

r' is fixed at .25. As the average between subtest inter-item correlation goes from .25 to zero, the differences between coefficient alpha and beta become quite apparent. Coefficient alpha remains very high and varies only slightly although the loadings of the items on the first factor change from .50 to zero and the correlations between halves of the test range from .87 to zero. Beta, on the other hand, is low when the between subtests correlation is low, moderate when it is moderate, and high when the test is truly unifactorial. It is also apparent from this example that beta is less sensitive to test length than is alpha.

Thus, in the case of a "lumpy test" (one with several large group factors) alpha overestimates the general factor saturation of the test and underestimates the total common factor saturation. Beta, on the other hand, gives a more appropriate estimate of the general factor saturation but severly underestimates the common factor saturation. Beta gives a better estimate of the test's homogeneity, while alpha is the more appropriate estimate of how well a test will correlate with another test sampled from the same domain.

Although beta does give a better indication of the lumpiness of a test than does coefficient alpha, it has at least one serious drawback when compared to alpha. Alpha can be found from the item and test variances without the inter-item covariance matrix. To find beta, on the other hand, requires finding the worst split halves of a test. To find this worst split half analytically requires trying all possible splits. For a test with twenty items, for example, and considering only splits of equal size, this requires an examination of 184,756 possible splits. Clearly an analytic solution for beta is impossible for any test of normal length (greater than ten to fifteen items). A simple heuristic, however, for estimating the worst split half is hierarchical clustering. But this is only a heuristic and will not always produce the worst split half. What is particularly interesting is that beta can be estimated by hierarchical clustering procedures and is also very useful as a stopping criterion in these very same clustering procedures.

Beta is a useful index in hierarchical clustering in that it can be used as a decision rule for combining two subtests into a higher order test. If the two subtests intercorrelate enough to produce a higher beta when they are combined than they have separately, then these two subtests should be considered to define a higher order test. If, however, the combined beta is less than the pooled estimate of beta, these subtests should not be combined, for the resulting test would have a smaller percentage of its variance associated with the general factor than do the two subtests.[1] For items sampled from one domain this rule will always result in subtests being combined, for as the number of equivalent items from a domain increases, beta will tend towards one. For items selected from two slightly related domains, this rule will prevent second order factors from emerging while the alpha rule will not.

To demonstrate this, consider two domains of items of size n with average within-domain correlation r' and average between-domain correlation of r. It is possible to show that if the unattenuated correlation between the two domains

$$[1] \qquad \tilde{r} = r/r' \geq [1 + (n-1)r'] \,/\, [n(2-r')]$$

then the alpha rule will allow these two domains to combine. The unattenuated correlation ($\tilde{r}$) must be

$$[2] \qquad \tilde{r} = r/r' \geq [1 + (n-1)r'] \,/\, [2 + (n-2)r']$$

for the beta rule to allow these two domains to combine. In the case of one of the examples, a twenty-item test with two ten-item subtests sampled from two different domains, this means that if the average within-domain correlation is .25, then the two subtests will be combined by the alpha rule if the average between-domain correlation is greater than .046, *i.e.,* if the unattenuated correlation between subtests is greater than .186. The beta criterion, on the other hand, would allow these two subtests to combine only if the average between-domain intercorrelation were greater than .203, *i.e.,* if the unattenuated correlation between subtests were greater than .81.

It can be seen from Equations [1] and [2] that the alpha rule becomes less stringent as the number of items in a cluster

[1] A more stringent decision rule would be to combine two subtests only if the combined beta is greater than the maximum of the two subtest betas. A less stringent decision rule would be to form one test if the combined beta is greater than the minimum of the two subtest betas.

increases, or as the average within-cluster inter-item correlation decreases. The beta rule, on the other hand, becomes more stringent as the number of items in the cluster increases and less stringent as the within-cluster inter-item correlation decreases. As clusters become larger, however, there is a normal tendency for the average within-cluster correlations to decrease. Sudden decreases in r thus will lead to the beta criterion not being met, while gradual decreases in r will satisfy the criterion. However, a difficulty with the beta rule exists in that it is possible for local minimum to exist. That is, when combining items that are initially highly correlated but that also have a large general factor, it is possible for beta to initially decline and then rise back to levels above the small subtest level.

Empirical Examples

The usefulness of cluster analysis using the beta criterion as a tool in scale construction can be shown in two ways. One is to compare cluster procedures to more conventional procedures such as factor analysis on artificial data, and the alternative is to demonstrate that it produces reasonable solutions on real data sets. Comparisons on artificial data sets have the advantage that the underlying structure is known but the disadvantage that they are in fact artificial. Solutions on real data sets are always open to the criticism that the "true" structure has not been found. Therefore, both comparisons will be made and three types of examples will be shown. All three will be comparisons of a hierarchical clustering algorithm using the coefficient beta criterion (ICLUST VI; Revelle, Note 3) to a commonly used "push button" factor analysis package available on SPSS (Nie et al, 1975). The first two comparisons are between cluster and factor analysis on artificial problems with oblique (Example 1) and orthogonal (Example 2) item domains. For both of these examples two replications of the cluster solution were done with four different sample sizes. In addition, for the comparison using orthogonal item domains, two different levels of item communalities were used. The final data set (Example 3) is a comparison of cluster analysis and factor analysis in forming scales from 92 items selected from a study of the common factors of the Guilford and Cattell inventories (Sells, Demaree, & Will, 1970). These 92 items were used as part of a project to develop a stress susceptibility scale (Revelle, Note 4).

For each of the examples both programs were used with their default values for an initial exploratory run. This always resulted in the formation of more factors than clusters. To allow for comparisons of solutions with the same number of scales both procedures were then used in a semi-confirmatory mode. In the case of factor analysis this meant rotating the first n factors, while in the case of cluster analysis this meant assigning items from extra clusters to the first n clusters on the basis of cluster loadings on these n clusters. To compare the adequacy of the solution, the number of items having their highest loading on the appropriate factor/cluster was found.

Example 1: Four Correlated Clusters

A 32-item population correlation matrix was made with a hierarchical structure similar to that found with 16 PF or EPI items. All items were given loadings of .32 on a general factor, one of two broad group factors and one of four narrow group factors. All other loadings were set to zero. Thus, there were four subsets of eight items each with average intercorrelations of .3; items in the first two of these subsets had average between subset correlations of .2, as did items within the last two subsets; and items within the first two subsets had average correlations with items from the second two subsets of .1. This structure is similar to that found in the EPI in which items in the Stability/Neuroticism scale have low correlations with the items from the Introversion/Extroversion scale, while within the I/E scale there are correlated sub-groups of items tapping sociability and impulsivity. Samples of size 50, 100, 200 and 400 simulated subjects were assigned scores on these 32 items.

To compare exploratory factor analysis with exploratory cluster analysis, both the SPSS factoring program and the ICLUST program were used with their default values and the number of clusters/factors obtained are reported in Table 2. In addition, to study the stability of clustering solutions, each cluster analysis was repeated on another sample of the same size. Finally, to compare the adequacy of the solutions, both the SPSS and ICLUST programs were used in a semi-confirmatory fashion, that is, four factor/cluster solutions were requested. When items were assigned to the factor/cluster on which they had their highest loading, it was possible to count how many items were correctly classified (*i.e.* that loaded on the correct group factor, Table 2).

Hierarchical cluster analysis using the beta criterion consistently identified fewer clusters (3-7) than did conventional factor analysis using the eigenvalue greater than 1.0 rule (6-11). When the correct number of factors to extract and rotate was specified, the scales formed by clustering were somewhat superior in the accuracy of classification of items to scales, particularly when the sample size was small. It is important to note that when these items were clustered using an increase in coefficient alpha stopping rule (Kulik, Revelle, & Kulik, Note 2) all of the items were formed into one large 32 item cluster for all of the data sets except the sample sizes of 50.

Thus, for the case of items with an oblique structure and low inter-item correlations, hierarchical clustering using the increase in coefficient beta stopping rule is an effective technique. But, since the factor rotation program (VARIMAX) used by SPSS as the default option was not meant to identify oblique factors, the comparison of a hierarchical procedure with one meant to perform best on non-hierarchical data is not completely fair. Therefore, a comparison of ICLUST using the beta criterion with "push button" factor analysis was done with a second data set, one with orthogonal factors.

Example 2: Four Orthogonal Clusters

A population correlation matrix with four factors was generated with average within cluster inter-item correlations of .3 (*i.e.,*

Table 2
Characteristics of Cluster and Factor Solutions: Oblique Case.

	Sample Size			
	50	100	200	400
Number of Clusters β^a	4-7	4	3-4	4-6
Number of Clusters α^b	2-5	1	1	1
Number of Factors $\lambda > 1.0$	10	11	9	6
Percent of Items Classified by 4 Clusters	69-91	97	100	100
Percent of Items Classified by 4 Factors	66	91	97	100

[a] The number of clusters identified using the beta criterion in both replications is shown.

[b] The number of clusters identified when using an increase in alpha criterion is shown.

factor loadings of .55) and all other correlations of zero. A second data set was generated with average within cluster inter-item correlations of .2. Once again, factor analysis and cluster analysis were compared on samples of size 50, 100, 200 and 400 simulated subjects with 32 items drawn from this population. The stability of the cluster solutions was studied by replicating each solution on a separate sample (Table 3). The conclusions are very similar to those drawn from example 1. Using the eigenvalue greater than 1.0 rule resulted in far too many factors being extracted (7-14), whereas using the beta criterion as a stopping rule for hierarchical clustering was much more accurate (4-9). When items were assigned to the first four rotated factors or largest four clusters, the number of items correctly assigned to factors/clusters was very close to perfect for both procedures and no systematic differences could be observed.

Table 3
Characteristics of Cluster and Factor Solutions: Orthogonal Case

	Communalities = .3			
	Sample Size			
	50	100	200	400
Number of Clusters	4-9	4-6	4-5	5
Number of Factors $\lambda > 1.0$	12	11	8	7
Percent of Items Classified				
by 4 Clusters	94	100	100	100
by 4 Factors	94	100	100	100
	Communalities = .2			
Number of Clusters	6	5-6	6-7	4-6
Number of Factors $\lambda > 1.0$	13	14	11	9
Percent of Items Classified				
by 4 Clusters	75-78	88-100	100	100
by 4 Factors	69	94	100	100

Thus, as in the oblique case, hierarchical clustering with the beta criterion proved to be very useful in determining the proper number of clusters to extract and in correctly classifying the items to the scales. Although these two examples show hierarchical clustering to be useful in forming scales with artificial data sets, it is also important to show utility with real data problems. The next example was chosen to represent a typical applied scale construction problem.

Example 3: Sociability, Impulsivity, Tension Items

As part of an ongoing research project studying individual differences in the response to stress, Revelle (Note 4) administered 92 items that included measures of sociability, impulsivity and nervous tension to 206 subjects. These items were taken from the study by Sells *et al* of the common factor structure of the Guilford and Cattell personality inventories. Previous studies (Revelle, Amaral, and Turriff, 1976; Gilliland, Note 1) have shown that some of these items were related to efficient performance under time pressure or caffeine-induced stress. Competing hypotheses about the nature of introversion-extroversion suggested that the sociability and impulsivity items either should (Eysenck, 1967) or should not (Guilford, 1975) form one scale.

ICLUST using the beta criterion identified thirteen clusters of which four each accounted for more than five percent of the total variance. When cluster solutions were found for three or four clusters, the first three clusters contained sociability, impulsivity, and nervous tension items, respectively. In the four cluster solution, seven items associated with a happy-go-lucky or carefree content were found to be separate from items with a sociability content. All clusters had substantial alpha reliabilities (.92, .79, and .80 for the three cluster and .91, .79, .80 and .74 for the four cluster solution) and adequate beta reliabilities.[2] They were only moderately intercorrelated (Table 4).

It is interesting to note that if the first two clusters were combined to form one scale, the content would be suggestive of Eysenck's introversion-extroversion dimension. This combined scale would have an alpha reliability of .91 which would normally be considered quite respectable for a scale of this length (53 items with an average intercorrelation of .15). However, coefficient beta for this combined scale would be only .44 which is less than the betas for either the sociability (.54) or the impulsivity (.51) scales. Thus, while coefficient alpha gives the impression that extroversion can be measured by a homogeneous scale, coefficient beta suggests that these two sub-components should not be combined.

Factor analysis found 29 factors with eigenvalues greater than 1.0. To allow for comparisons with the cluster solution, the

[2] Since beta estimates the first factor saturation of a test, one might want to have a beta value greater than .50. This would be equivalent to the requirement that at least 50 percent of a test's variance is associated with the first factor of that test.

Table 4
Final Cluster Solution to 92 Sociability, Impulsivity and Tension Items.

Cluster	Number of Items	α	β	Representative Items
I	30	.91	.53	Likes to mix socially with people Easy to make new acquaintances Difficulty making new friends (minus)
II	20	.79	.51	Inclined to be quick in actions Rates self as impulsive person Often feels bubbling with energy Rushes from one activity to another
III	20	.80	.53	Over-excited and rattled in upsetting situations Rates self as tense individual Becomes irritated over little annoyances
IV	7	.74	.45	Rates self as a happy-go-lucky individual Ordinarily a carefree individual Is inclined to be over conscientious (neg.)

Intercorrelations

	3 Cluster Solution			4 Cluster Solution			
	I	II	III	I	II	III	IV
I	(92)			(.91)			
II	.32	(.79)		.30	(.79)		
III	–.17	.16	(.80)	–.13	.15	(.80)	
IV				.40	.21	–.18	(.74)

largest four factors were rotated to a Varimax criterion. Four scales then were formed by finding the sum of the unit-weighted items for all items with loadings greater than .3. The resulting reliabilities were .91, .74, .80, and .79. These four factor scales had slightly higher average absolute inter-correlations than did the four cluster scales. When the factor scales were purified by assigning items to only one scale, and using all items with loadings greater than .25 (this is similar to the purification done for the clusters), the reliabilities were reduced slightly as were the scale intercorrelations.

Substantively, the first and third factors were very similar in content to the first and third clusters (all 30 items in the first cluster were included in the 33 highest item loadings on the first factor; similarly 18 of 20 items in the third cluster were included among the 21 best items on the third factor). The impulsivity items in the second cluster and the carefreeness items from the fourth cluster were assigned to the second factor, while some activity items from the second cluster had the highest loadings on the fourth factor.

Discussion and Recommendations

All three of the examples had the low inter-item correlations typical of many personality inventories (*e.g.,* Sells et al., 1970). The two simulated problems indicated that with such data, "push-button" factor analysis overestimates the number of factors to extract. The final example suggested that this problem occurs with real data as well. Cluster analysis as exemplified by the ICLUST algorithm was not as susceptible to overfactoring. When the proper number of factors was specified on the artificial problems, both factor analysis and cluster analysis were equally able to classify items correctly. In the final example, with real data, far fewer clusters were identified than were factors. When both procedures were requested to produce four cluster (factor) solutions, the solutions were quite similar.

It is important to point out, however, that the comparisons with factor analysis were done using default values. This is the kind of analysis typical of the naive factor analysis user. It is likely that sophisticated analysts would have achieved solutions equivalent to the default cluster solutions had they carefully compared alternative solutions and used the intuitive skills that come from years of experience at examining factor outputs.[3]

From these examples the following tentative recommendations can be made to the investigator interested in forming composite scales from batteries of items.

1) The number of scales or indices to be formed from a set of items should not be determined solely by the conventional factor analytic procedure of extracting all factors with eigenvalues greater than 1.0. Rather, the number of scales to form should be determined by some method that tests for the psychometric adequacy of each scale.
2) The adequacy of a scale as a measure of a single construct should not be assessed solely by the magnitude of coefficient alpha or the average loadings of items on the scale, but also by the magnitude of the worst split-half reliability coefficient beta.
3) When forming scales from sets of items, hierarchical clus-

[3] In all fairness to factor analysis, it should be pointed out that some of the most experienced practitioners of factor analysis do not encourage the use of factor analysis on items, but suggest that parcels (Cattell, 1973) or Factorially Homogeneous Item Dimensions (FHIDS; Comrey, 1961) should be formed first and that these then should be factored. These recommendations are excellent but unfortunately are rarely followed.

tering procedures using the increase in coefficient beta stopping criterion can be particularly useful.

Conclusions

Standard practices in scale construction involve unit weighting of items with high loadings on the same factor and then testing for internal consistency by finding coefficient alpha of the resulting test. The number of factors to extract depends upon the experimenter's theory, sophistication, and guesses. A test composed of items with moderate loadings on the first principal component can have a high coefficient alpha and yet still represent several distinct constructs. One way to test for this condition is to find the worst split-half reliability (coefficient beta). If a test has a sizable beta as well as a sizable alpha, the test can be considered to be assessing one construct. A high alpha and a low beta, on the other hand, is an indication that the test is "lumpy" and has several large group factors. Such a test should not be considered to be a measure of one construct, but rather of two or more.

An advantage of coefficient alpha is that it can be found without finding the inter-item correlation matrix. To find coefficient beta, on the other hand, requires an analysis of the inter-item correlations. Any procedure that partitions a test into two sub-tests such that the within sub-test variance is maximized (and the resulting between sub-test covariance is minimized) will give coefficient beta of the test. A simple algorithm for estimating coefficient beta uses principles of hierarchical cluster analysis. The advantage of this hierarchical procedure is that it not only estimates beta for the entire test, but for the sub-tests as well. A test with a lower beta than the betas associated with its sub-tests should not be considered a very good test. One that has a higher beta for the total test than for the sub-tests is a better test.

Hierarchical clustering procedures are most appropriate when the variables to be clustered have some hierarchical structure (*i.e.*, a general factor, several common factors, and then several specific factors). However, the procedure also is suitable for the case when variables can be partitioned into truly orthogonal components with a simple structure solution.

A final advantage of hierarchical procedures is that they are very simple to understand and economical to perform.[4] Although

[4] Computer costs for cluster analysis are between 10 and 50 percent of that needed for a simple factor analysis.

simplicity is not normally considered a virtue, perhaps it should be. The advent of high speed computers and powerful statistical packages makes very sophisticated methodologies available to the most naive user (Kaiser's "sweet young thing"). This often results in ill-conceived studies seeming impressive because they have made use of complicated analytical procedures. Until computer packages such as BMD or SPSS will accept commands only from users who have passed an interactive test of their psychometric knowledge and ability, or until there are levels of programs available for various levels of user ability, there is a great benefit in using scaling algorithms that are simple to understand and robust to violations of their assumptions. Hierarchical clustering procedures can claim to be both.

Finally, it should be pointed out that as with all other scaling procedures, cluster analysis cannot and will not replace common sense and carefulness. The investigator who hopes that he or she can administer a battery of sloppily written items and discover some gem of truth in the resulting clutter of clusters is mistaken. The one thing he or she will discover is that if the items are bad to start with, the resulting cluster-scales will have low estimates of internal consistency.

Reference Notes

1. Gilliland, K. The interactive effect of introversion-extroversion with caffeine induced arousal on verbal performance. Unpublished doctoral dissertation, Northwestern University, 1976.

2. Kulik, J. A., Revelle, W., & Kulik, C. L. C. Scale construction by hierarchical cluster analysis. Unpublished paper, University of Michigan, 1970.

3. Revelle, W. *ICLUST: A program for hierarchical cluster analysis.* Northwestern University. Computing Center Document 482, 1977.

4. Revelle, W. Development of a stress susceptibility scale. Unpublished paper, Northwestern University, 1978.

References

Blashfield, R. K. Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, 1976, *83*, 377-388.

Cattell, R. B. *Personality and mood by questionnaire.* San Francisco: Jossey-Bass, 1973.

Comrey, A. Factored homogeneous item dimensions in personality research. *Educational and Psychological Measurement*, 1961, *21*, 417-431.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, *16*, 297-334.

Everitt, B. *Cluster Analysis.* New York: Wiley, 1974.

Eysenck, H. J. *The biolgical basis of personality.* Springfield: C. C. Thomas, 1967.

Goldberg, L. Parameters of personality inventory construction and utilization: A comparison of prediction strategies and tactics. *Multivariate Behavioral Research Monographs,* 1972, Number 2.

Guilford, J. P. *Psychometric Methods.* New York: McGraw-Hill, 1954.

Guilford, J. P. Factors and factors of personality. *Psychological Bulletin,* 1975, *82,* 802-814.

Hartigan, J. A. *Clustering Algorithms.* New York: Wiley, 1975.

Loevinger, J., Gleser, G. C. & DuBois, P. H. Maximizing the discriminating power of a multiple score test. *Psychometrika,* 1953, *18,* 309-317.

McQuitty, L. L. & Koch, V. L. Highest entry hierarchical clustering. *Educational and Psychological Measurement,* 1975, *35,* 751-766.

Nie, N. H., Hull, C. H., Jenkins, J. G. Steinbrenner, K. & Bent, D. H. *Statistical package for the social sciences (2nd edition).* New York: McGraw-Hill, 1975.

Nunnally, J. *Psychometric Theory.* New York: McGraw-Hill, 1967.

Revelle, W., Amaral, P., and Turriff, S. Introversion/extroversion, time stress, and caffeine: the effect on verbal performance. *Science,* 1976, *192,* 149-150.

Sells, S. B., Demaree, R. G. and Will, D. P., Jr. Dimensions of personality: I. Conjoint factor structure of Guilford and Cattell trait markers. *Multivariate Behavioral Research,* 1970, *5,* 391-422.

Tryon, R. C. & Bailey, D. E. *Cluster analysis.* New York: McGraw-Hill, 1970.

Wiggins, J. S. *Personality and prediction: Principles of personality assessment.* New York: Addison Wesley, 1973.

Wishart, D. An algorithm for hierarchical classifications. *Biometrics.* 1969, *25,*(1), 165-170.

APPENDIX

Consider two subtests of length n with average within-test inter-item correlation of r' and average between-test inter-item correlation of r. Then

$$\alpha_a = \alpha_b = nr'/(1+(n-1)r')$$

and

$$\alpha_{ab} = \frac{2n^2r+2n(n-1)r'}{2n^2r+2n(n-1)r'+2n} \cdot (2n)/(2n-1) \quad .$$

The two subtests should be combined if $\alpha_a<\alpha_{ab}$, *i.e.,* if

$$\frac{(2n-1)r'}{1+nr'-r'} < \frac{2(nr+nr'-r')}{nr+nr'-r'+1} \, .$$

Multiplying and collecting terms reduces this to the expression

$$\alpha_a<\alpha_{ab}<=>r'(nr'-r'+1)<nr(2-r')$$

or

$$\frac{1+(n-1)r'}{n(2-r')} < r/r' \quad .$$

Similarly, for subtests of length $n = 2m$,

$$\beta_a = \frac{4m^2r'}{2m^2r'+2m(m-1)r'+2m} \text{ and}$$

$$\beta_{ab} = \frac{16m^2r}{8m^2r+4m^2r'+4m(m-1)r'+4m} \quad .$$

The two subtests should be combined if $\beta_a < \beta_{ab}$, *i.e.*, if

$$\frac{r'}{2(mr'+(m-1)r'+1)} < \frac{r}{2mr+mr'+(m-1)r'+1}$$

Multiplying and collecting terms reduces this expression to the following:

$$\beta_a < \beta_{ab} <=> \frac{1+2mr'-r'}{2(1+mr'-r')} < r/r'$$

substituting n for $2m$, this becomes

$$\beta_a < \beta_{ab} <=> \frac{1+(n-1)r'}{2+(n-2)r'} < r/r' \quad .$$