Behavior Research Methods & Instrumentation
1978, Vol. 10 (5), 739-742

ICLUST: A cluster analytic approach to exploratory and confirmatory scale construction

WILLIAM REVELLE
Northwestern University, Evanston, Illinois 60201

A common problem in the social sciences is to form a set of relatively independent and internally consistent scales from a large pool of items. Frequently, these scales are formed by simply summing the responses to keyed items. The problem, then, is to determine how best to partition the initial set of items into subsets or scales that are highly internally consistent and relatively independent. A common alternative is to factor analyze the interitem correlation matrix and then to select items on the basis of factor loadings. Those items with a high loading on a particular factor are combined into a scale by applying unit weights to the items. This method, although probably the most common scale construction procedure, has several drawbacks: Interitem correlations are usually small (average interitem correlations $\leqslant .3$) and the sample sizes are usually not much larger than the number of items. These problems tend to lead to overfactoring (extracting too many factors), unstable rotations, and generally nonsensical solutions. In fact, because of the problems encountered in factoring items, many experienced factor analysts recommend against such procedures (Cattell, 1973; Comrey, 1961; Nunnally, 1967). However, a sampling of journals in the social sciences suggests that this advice is rarely followed.

When the item pool is large (greater than 10-20 items), when the item intercorrelations are small (between 0.0 and .5), or when the sample sizes are small, an alternative method that is particularly appropriate is cluster analysis. Cluster analysis is a loosely defined set of procedures associated with the partitioning of a set of objects into nonoverlapping groups or clusters (Everitt, 1974; Hartigan, 1975). Although normally used to group objects, occasionally cluster analysis has been applied to the problem of grouping variables and, as such, is similar to procedures of group factor analysis (Loevinger, Gleser, & Dubois, 1953; Revelle, in press; Tryon & Bailey, 1970). A disadvantage for scale construction of many clustering procedures is that they do not include basic psychometric decision rules to evaluate either the quality or the number of clusters to extract. It is possible, though, to combine psychometric principles with clustering procedures. This combination results in a simple but useful approach to scale construction, and, for forming scales from items, may be compared favorably with more elegant factoring algorithms. Clustering's advantage over more complex factoring algorithms (principal factor, maximum likelihood, etc.) is that clustering is specifically designed for the task at hand: finding unit-weighted item composites that are internally consistent and independent of other composites. This is, of course, also the goal of factoring and rotation procedures; and clustering and factoring normally give similar solutions when the same number of clusters (factors) is extracted (Revelle, in press).

The ICLUST (item cluster analysis) package of programs is a group of routines for performing exploratory and confirmatory scale construction. The exploratory routines use a hierarchical clustering algorithm to determine an initial grouping of the variables into clusters based upon the principle that scales should be combined into higher order scales if and only if the internal consistency of a higher order scale is greater than that of its component scales. The scales identified through this initial routine may then be subjected to a step-down iteration procedure that compares the overall quality of cluster solutions using different numbers of clusters. Alternative cluster solutions, and their fit to the very simple structure (VSS) criterion (see Revelle, Note 1), are presented for user inspection. An alternative use of the ICLUST package is to do confirmatory cluster analysis, which involves defining certain sets of items as cluster scales and then examining the internal structure of each of these scales, as well as the quality of the overall solution. Yet a third possibility is to do a mixed confirmatory-exploratory analysis, in which certain items are forced by the user to form scales and other items are assigned by the program. Once again, statistics of clustering adequacy are reported for this option as well.

ICLUST Algorithms. The ICLUST package makes use of two different clustering algorithms. The first is a hierarchical algorithm that is used for the initial definition of clusters and for the evaluation of the quality of individual clusters. This algorithm is used in exploratory analyses to determine the number and membership of unpurified clusters, and in confirmatory analyses to determine the internal consistency of the final clusters. The second algorithm is a nonhierarchical one that is used for cluster purification, confirmation, and for step-down iterations.

The hierarchical clustering algorithm may be summarized as follows: (1) Find the interitem proximity matrix. (2) Find the most similar pair of variables in this matrix. (3) If the internal consistency of the cluster formed by combining this pair of variables would be greater than that of its two components, then combine the two variables into a new (composite) variable. If this test is not satisfied, return to Step 2 and find the next most similar pair of variables. (4) If the test at Step 3 is passed, add the new composite

Correspondence concerning this article should be addressed to William Revelle, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

variable to the set of previous variables, delete its two component variables, and calculate the proximity of the new composite with the remaining variables. (5) Repeat Steps 2-4 until no more variables pass the increase-in-internal-consistency criterion of Step 3. (6) Find the value of the VSS goodness-of-fit criterion.

The proximity matrix found in Step 1 is the matrix of Pearson correlations. In the case of dichotomous items, this is equivalent to the phi coefficient. As a user-controlled option, either the raw correlation or a correlation corrected for cluster unreliability may be used as the proximity measure. The reliability estimate is the highest correlation an item has; cluster reliability is estimated using coefficient alpha (Cronbach, 1951). The effect of using the highest correlation as a reliability estimate is to form the initial clusters from reciprocal pairs of variables, that is, variables that have their highest correlations with each other (McQuitty & Koch, 1975).

In Step 2, in order to insure that the clusters are reasonably compact, and in order to make the searching routine faster, a list is formed of the most similar variable for each cluster variable. This list is then searched for its maximum value and the program proceeds to Step 3. When variables are combined at Step 4, this maximum value list is updated. By limiting the search to those pairs of variables in which one member or the other is most similar to the other member of the pair, it is possible to avoid clusters being formed from variables that have higher proximities with other variables but fail to meet the increase-in-internal-consistency criterion. That is, if variables x and y are each more similar to variable z than they are to each other, but neither x nor y can combine with z because of the internal consistency criterion, then x and y will not be allowed to combine with each other.

Step 3 is the most important element in the ICLUST exploratory routine. Two variables (or clusters) are formed into a higher order cluster only if both of two tests of internal consistency are passed. The first test is that coefficient alpha (Cronbach, 1951) of the composite should be greater than that of either of its two components. This test is very appropriate when single items are combined into a large cluster, and its use has been suggested previously by Loevinger et al. (1953) for the case of nonhierarchical clustering and by Kulik, Revelle, and Kulik (Note 2) for the case of hierarchical clustering. Unfortunately, for the case of hierarchical clustering, the criterion is not very useful for testing whether large clusters should be combined. It has been shown (Revelle, in press) that, as the cluster size increases, almost any two clusters will meet the increase-in-coefficient-alpha criterion.

A more appropriate test for hierarchical clustering is the application of a test for an increase in coefficient beta. Beta is defined to be the worst split-half reliability of a test (Revelle, in press). If coefficient beta of the composite cluster is greater than the average beta of the two components, then these two components should be combined. If, on the other hand, beta of the composite would be less than the average beta of the components, then the composite would be less homogeneous than these components and should not be formed. When compared to coefficient alpha, the use of coefficient beta always is more conservative and becomes even more conservative as cluster size increases, but relatively less conservative as cluster homogeneity decreases (Revelle, in press).

To better understand the relationship between these two coefficients of internal consistency, consider the following example. Consider a test formed by combining two unrelated subtests which themselves are internally consistent. Assume each subtest has 10 items. Let the average interitem correlation within each subtest be .25 and the average interitem correlation between the two subtests be equal to 0.0. This means that each subtest has an alpha of .77 and an average item-to-whole correlation of .44. Since the two subtests are unrelated, they should not be considered to form one test. But the conventional estimates of internal consistency for such a test would be high. In this example, alpha for the entire test would be .73, and the average item-to-whole correlation would be .31. These values are typical for tests of such length. Coefficient beta, on the other hand, being based upon the correlation between the two worst halves of the total test, would properly reflect that the total test is made up of unrelated parts. In this example, coefficient beta would be 0.0. In the case that a test is truly univocal, alpha and beta will give similar estimates of internal consistency, although beta, being based upon the worst split half, will always be less than or equal to alpha. The "lumpier" a test, the greater will be the disparity between alpha and beta.

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 is independent of the order in which items are combined. Exact calculation of beta, on the other hand, is dependent upon finding the worst split half of a test. To find the worst split half analytically requires considering all possible splits. For the 20-item example, and considering only splits of equal size, this requires examining 184,756 possible splits. Beta can be estimated, however, by using hierarchical clustering procedures (Revelle, in press). Thus, beta can be estimated by hierarchical clustering procedures and also can be used by these same procedures as a stopping criterion.

The application of the alpha and beta criteria for forming higher order clusters allows for a dynamic stopping criterion. Rather than stop clustering when some arbitrary value of homogeneity is passed, ICLUST will form clusters as long as the higher order clusters are more internally consistent than their components.

When they would not be, they should not be formed into clusters, for further combination would obscure and reduce their interpretability.

The fourth step of the hierarchical algorithm involves calculating the proximity of the new composite cluster with the remaining clusters or variables. This is done by standard psychometric principles. That is, the correlation of two tests is the sum of their unweighted interitem covariances divided by the square root of the product of their variances.

Steps 2-4 are repeated until no new clusters pass the increase-in-internal-consistency criteria of Step 3. At this point, an overall summary statistic of the goodness of fit of the entire solution is calculated. This is the VSS criterion, which measures how well the cluster solution reproduces the initial correlation matrix (Revelle, Note 1). To find the VSS criterion, a predicted correlation matrix ($\hat{R}$) is formed according to the cluster analytic equivalent to the general factor equation:

$$\hat{r}_{ij} = r_{ic_i} r_{ji_j} r_{c_i c_j}. \qquad (1)$$

That is, the predicted correlation between the ith and jth items ($\hat{r}_{ij}$) is the product of the loading of the ith item on the cluster with which it has its highest loading (r_{ic_i}), the loading of the jth item on its defining cluster (r_{jc_j}), and the intercorrelation between these two clusters ($r_{c_j c_j}$). This special case of the general factor law is formed by assuming that each item is of Rank 1 and that, therefore, all loadings other than the greatest loading are zero. To the extent that $\hat{R}$ is a good fit to R (the observed correlation matrix), the clustering model is appropriate. The goodness-of-fit index is formed by finding the mean squared residuals,

$$MS_{r*} = \frac{\sum\sum_{i<j} r^{*2}_{ij}}{df} = \frac{\sum\sum_{i<j} (r_{ij} - \hat{r}_{ij})^2}{df}, \qquad (2)$$

and comparing them to the mean of the original squared correlations.[1] This ratio is then subtracted from 1 to give an index of fit:

$$VSS_1 = 1 - \frac{MS_{r*}}{MS_r}. \qquad (3)$$

It is important to note that, if a true cluster solution exists [i.e., if each item is of Complexity 1 but is embedded in a matrix of higher rank (K)], then the index will be maximized if K clusters are extracted (see Revelle & Rocklin, Note 3, for examples). Second, if the cluster solution is rotated away from the simple true structure, then the goodness of fit will also diminish. That the goodness-of-fit test peaks at the appropriate number of clusters is particularly useful for evaluating the relative quality of various solutions. This index is a psychometric goodness of fit, and it should not be used for determining significance tests. The distribution of the residual correlations found by using Equation 2 is not known.

A problem with hierarchical clustering for scale construction is that it is possible for items to be grouped into clusters with which they do not have their highest correlations.[2] To avoid this unfortunate consequence of hierarchical clustering, ICLUST derives the initial cluster solution using a hierarchical algorithm, but this solution is then purified by reassigning items that have been misclassified. The cluster purification algorithm may be summarized as follows: (1) Identify the cluster centroids. (2) Calculate item by cluster correlations. (3) Assign items to the cluster with which they correlate most highly. (4) Return to Step 1 until no more items are reassigned or until a certain number of iterations have been done.

The cluster centroids found in Step 1 may be either those identified by the hierarchical routine (for an exploratory analysis) or those prespecified by the user (for a confirmatory analysis). For confirmatory runs used to evaluate the quality of a particular a priori solution, the initial centroids are formed from the a priori scales.

In the exploratory mode, after clusters have been determined by the initial hierarchical procedure with purification iterations, the quality of the overall solution is assessed by means of the VSS criterion. The initial solution may then be "stepped down" to progressively fewer clusters by repeated use of the cluster purification cycle. At each step-down level, the cluster from the preceding level that accounted for the least variance is discarded, and items assigned to that cluster are reassigned to the remaining clusters. Also, at each step-down level, the values of alpha and beta for each cluster and the goodness-of-fit index (VSS) is calculated. This allows the user to compare the quality of various solutions, in order to determine which one to consider final. Step downs are not automatic, but they may be requested.

Statistics Reported. Three types of statistics are reported for each analysis: those having to do with the characteristics of the overall solution, those having to do with the quality of particular scales, and those having to do with individual items.

The best description of the overall quality of a cluster solution is the VSS criterion. VSS values for both orthogonal and oblique clusters are reported. The "orthogonal" VSS is an index of how well the solution fits when the between-clusters correlations in Equation 1 are set to 0.0 for items not defining the same cluster. The mean squared residual correlations for both the orthogonal and oblique solutions are reported, as is the mean square of the original correlation matrix.

The quality of each particular scale can be evaluated by the value of coefficients alpha and beta, as well as the average interitem correlation within the cluster, and the percentage of total variance for which a cluster accounts. The intercorrelations of the clusters can be used as additional indications of which clusters are most independent of the remaining clusters.

Statistics reported for individual items include the mean, variance, minimum, and maximum values, as well as item-cluster correlations (loadings). The cluster "loadings" are corrected for item-whole overlap and for cluster unreliability. Uncorrected correlations are also reported.

The interpretability of each cluster and the relation of the items to the clusters is shown in a summary table in which each item is listed in descending order of its (absolute) correlation with its defining cluster. As an additional aid to interpretation, up to 75 characters of identification (i.e., the content of the item) are listed for each item in this summary table.[3]

Availability and Cost of Operation. ICLUST was written in FORTRAN IV for a CDC 6400-6600 series computer with extended core storage (ECS). It has been adapted to other CDC systems without ECS and to IBM 370 equipment.[4] On a CDC 6600, it takes approximately 10 sec to find a purified solution for 57 variables, 30 sec for 92, 50 sec for 140, and less than 300 sec for 290 variables. The current compilation is limited to 300 variables with no limit on subjects. To facilitate semi-interactive use, ICLUST saves the initial correlation matrix, which can be used repeatedly for later restarts comparing different solutions. This allows the user to examine the output from an exploratory run, decide how many clusters to retain, and then proceed to do step-down analyses.

REFERENCE NOTES

1. Revelle, W. *Very simple structure: An alternative criterion for factor analysis.* Paper presented at the annual meeting of the Society for Multivariate Experimental Psychology, Colorado Springs, Colorado, November 1977.
2. Kulik, J. A., Revelle, W., & Kulik, C. L. C. *Scale construction by hierarchical cluster analysis.* Unpublished manuscript, University of Michigan, 1970.
3. Revelle, W., & Rocklin, T. *Alternative procedures for estimating the optimal number of interpretable factors.* Paper presented at the European meeting of the Psychometric Society, Uppsala, Sweden, 1978.
4. Revelle, W. *ICLUST: A program for analyzing the internal structure of tests.* Northwestern University Computing Center Document No. 432, 1977.

REFERENCES

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.

Hartigan, J. S. *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.

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

Revelle, W. Hierarchical cluster analysis and the internal structure of tests. *Multivariate Behavioral Research,* in press.

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

NOTES

[1] The degrees of freedom are taken to be one less than the number of correlations [n(n – 1)/2] minus the number of intercluster correlations.

[2] Consider eight variables arrayed on a line ranging from 0 to 100. Applying a hierarchical clustering algorithm using either centroids or diameters to assess distance produces the following two-cluster solution:

((0 13) (31 (46 60))) (79 (90 100)).

But 60, although included in Cluster 1 (0, 13, 31, 46, 60), is actually closer to the centroid (89.67) of Cluster 2 (79, 90, 100) than it is to the centroid of Cluster 1 (30). Similarly, applying the criterion of cluster diameters, 60 is closest to Cluster 2, although hierarchical analysis assigned it to Cluster 1. In actual analyses, between 5% and 15% of the items are misclassified according to this criterion.

[3] For a more detailed listing of the user options available, the statistics reported, and procedures for using the program, consult the ICLUST users' guide (Revelle, Note 4).

[4] To obtain a users' manual, program listing, sample runs, and a computer tape with the object deck and test data, send $25 to William Revelle, Department of Psychology, Northwestern University, Evanston, Illinois 60201.

(Accepted for publication June 1, 1978.)