Gorsuch R. L. (2003). Factor Analysis. In J. A. Schinka & W. F. Velicer (Eds.), Handbook of psychology: Volume 2 research methods in psychology (pp. 143-164). Hoboken, NJ:John Wiley & Sons.

CHAPTER 6

Factor Analysis

RICHARD L. GORSUCH

A major task of science is to develop theoretical constructs that bring together many observed phenomena. Historical examples of doing this include both ability and personality research. In the former, the moderate to high correlations observed among ability measures have led to the theoretical construct of general intelligence. In the latter, the moderate to high correlations among personality variables such as emotionality and frustration have led to the theoretical construct of anxiety (also called neuroticism). The construct validity of these theoretical constructs has been examined by factor analyses. Factor analysis is a statistical technique that reproduces the data by as few factors (potential theoretical constructs or latent variables) as possible.

A popular current use for factor analysis is scale development. When selecting a subset of the items for a scale, one needs to know how many constructs might be measured from the item pool and which items could measure each construct. This information is provided by a factor analysis. The items are factor analyzed to find the fewest number of factors that can represent the areas covered by the items. The relationship of each item to the factors indicates how it might be used in measuring one of the factors.

Whereas a factor analysis might result in a scale to measure a theoretical construct in a future study, confirmatory factor analysis and extension analysis in exploratory factor analysis allow another option. Factor analysis can be used in a new study to confirm or disconfirm the relationships between factors themselves or with other variables not in the factor analysis. No sales or factor scores are needed.

Although no factor analysis is ever completely exploratory—there is always an underlying theoretical model by which the data are collected—some factor analyses are primarily exploratory with no hypotheses, and others are primarily confirmatory, specifically testing hypotheses. Both types of factor analysis are examined in this chapter.

The purpose of this chapter is to provide a basic but comprehensive treatment of factor analysis. The intention is to give the reader the background to read, appreciate, and critique research from a factor analytic perspective, whether it be an article using factor analysis, an article using factor analysis inappropriately, or an article that could be strengthened if factor analysis were used. While no particular statistical package is assumed, this chapter also provides material needed to select the options for a factor analysis that are most appropriate to the purpose of the study.

The chapter starts with the basic equations and definitions of factor analysis. This section introduces the terms needed to understand factor analytic models and variations in the models. The second section of the chapter presents factor models, including component analysis (CA) and common factor analysis (CFA). CFA includes both exploratory (ECFA) and confirmatory (CCFA) factor analysis. In addition, all of these variants can be used with correlated or uncorrelated factor models. Presented with each model is the essential theoretical information to understand the model and the essential practical information to use the model.

Rather than reviewing all the possible procedures that could apply to each model, each section includes the

procedures that now have sufficient empirical and theoretical support to be the generally desired procedures for that model. In some cases, however, there are still minor variations in what procedure is used, and these are discussed with the model to which the variations apply.

Although the last decades have led to clear choices of some procedures over others for one or more models, several areas in factor analysis still present major unsolved problems. Three such problems are addressed after the models have been presented. The first is the continuing debate between advocates of two types of exploratory analysis: components and common factor. Second is the issue of how many factors to extract from a particular data set. Third is the question of how the factors in one data set can be related to other variables that were in the data set but were not included in the factor analysis, and how factors may be related across studies.

The concluding section points to elements of all good research designs that need to be remembered in designing a factor analytic study. Included in this section are discussions of the need for high-quality variables and how many cases are needed.

Three examples are used to illustrate factor analysis. The first example is of six psychological tests for which the structure is easily seen in the correlation matrix (Gorsuch, 1983). Three of the variables are related to verbal ability and three to anxiety. The second example is a case in which we know what the factors should be: boxes (Gorsuch, 1983). Graduate students took 10 measures from ordinary boxes they found in their homes. Because these are all measures within three-dimensional space, we expect the factors to be those three dimensions: length, height, and width.

The third example uses the Canadian normative sample for the Wechsler Adult Intelligence Scale–III (WAIS-III; Gorsuch, 2000). The published correlation matrix among the scaled scores form the basis of analysis. The factor structure of the WAIS, and its children's version, the WISC (Wechsler Intelligence Scale for Children), have been extensively analyzed. (Detailed discussions of factor analytic topics are in Gorsuch, 1983; when no other references are provided, please consult that reference.)

BASICS OF FACTOR ANALYSIS

The purpose of factor analysis is to parsimoniously summarize the relationships among that which is being factored, referred to here as variables, with a set of fewer constructs, the factors. The analysis serves as an aid to theory development and scale construction. The term *variables* is used because most factor analyses are of scales and measures to which that term is immediately applicable; however, other types of data, such as people, can be used (see Gorsuch, 1983; Thompson, 2000).

Understanding is aided when several variables are found to correlate sufficiently so that they are measuring the same construct (i.e., factor). In the area of intelligence, for example, scales with labels of *vocabulary* and *similarities* correlate highly together and can be considered manifestations of verbal ability. Because vocabulary and similarities have been found to relate to the same factor, theoretical development may account for vocabulary and similarities simultaneously by accounting for the factor.

Scale construction is aided when the correlations among items show the items to fall into a certain number of clusters or groups. In psychology of religion, motivation items, for example, fall into groups of items representing an intrinsic motivation (e.g., *the main reason I go to church is to worship God*) and extrinsic motivations (e.g., *the only reason to go to church is to meet friends*). The items fall into several groups so that within a group the items correlate with one factor and not with the other factors. Items can then be picked by their correlations with the factors to form scales.

Note that there is little generalization across factors (because the variables of one factor do not correlate with the variables of another factor) and so factor analysis identifies qualitatively different dimensions. Within a factor there is generalization identified with quantitative differences (i.e., how each variable correlates with the factor).

In addition to the classical factor analysis of scales, there are other uses of factor analysis. It can be used to reduce several problems encountered in data analysis.

One problem in data analysis is the *multiple collinearity* problem. This occurs when several scales that are designed to measure the same construct are used in the same study. Such scales correlate so well that it affects the statistics, such as multiple correlation. First, with multiple collinearity, multiple regression beta weights are unstable, and therefore are difficult to replicate. Second, another degree of freedom is used for each additional scale that measures what one of the other scales also measures. Yet having the additional measures is desirable because they increase the overall accuracy of the study. Multiple collinearity can be among either the independent or dependent variables.

A solution to the multiple collinearity problem is to factor the variables; then the factors are used instead of the variables. The same domains are covered with the factor analysis as the ones covered by the variables, and the factor analysis also shows the overlap among the scales. The multiple collinearity among the factors will be low.

Another problem with statistics such as multiple correlation is that the regression weights have all the covariation

among the variables eliminated. It does this by partialing out the other variables from the weights. The common—that is, predictive variance that two or more variables have in common—may not be seen at all in the beta weights. Hence, a multiple regression can be significant even though none of the weights are significant; it is the variance that the variables have in common that predicts the dependent variable. The solution is to extract as many factors as there are variables and restrict the solution so that the factors are uncorrelated. These are then orthogonalized versions of the original variables. When these are used as the predictors in a multiple regression, all of the covariation is distributed among the variables and appears in the weights.

Development of factor analysis as a statistical procedure proceeds from the generalized least squares (GLS) model used in regression and other least squares analyses. Assuming all variables to be in Z score form for convenience, the model is based on this set of equations:

$$\begin{aligned}
X_{i1} &= w_{1A}A_i + w_{1B}B_i + w_{1C}C_i + w_{1D}D_i + \cdots + u_{i1} \\
X_{i2} &= w_{2A}A_i + w_{2B}B_i + w_{2C}C_i + w_{2D}D_2 + \cdots + u_{i2} \\
X_{i3} &= w_{3A}A_i + w_{3B}B_i + w_{3C}C_i + w_{3D}D_i + \cdots + u_{i3} \\
&\cdots \\
X_{iv} &= w_{vA}A_i + w_{vB}B_i + w_{vC}C_i + w_{vD}D_i + \cdots + u_{iv}
\end{aligned} \tag{6.1}$$

where, for the first line, X is the score for person i on variable 1, w is the weight for variable 1 for factor A, and A is the score for person i on factor A. The equation shows factors A through D and indicates that there may be more. Additional variables are indicated, for a total of v variables in the analysis.

The last element of each equation, u, is that which is unique to that particular variable, often called *error* or *residual.* Each u is in a separate column to indicate that each is distinct from any other u. There are as many distinct us as there are variables. It is important to note that each variable's uniqueness (us) includes two sources of variance. First is random error due to unreliability and second is that variance in the variable that is not estimable from the factors.

When the preceding equation is solved for each dependent variable, the multiple correlation of the factors with that variable can be computed. In factor analysis, the square of that multiple correlation is called the *communality* (h^2) because it is an index of how much that variable has in common with the factors.

How high can the communality be? The absolute maximum is 1.0, because then all the variation of the variable would be reproduced by the factor. But the psychometric maximum is the variable's reliability coefficient, which by definition is the maximum proportion of the variable that can be reproduced from a perfect parallel form, although occasional capitalization on chance may produce a sample communality slightly above the reliability. (Note: The reliability referred to in this chapter is always the reliability in the sample for the factor analytic study.) Of course, the reliability gives the communality only if all the nonerror variance is reproduced by the factors. The more likely result is that the factors reproduce only part of the reliable variance, and so the communalities are expected to be less than the reliabilities.

While Equation 6.1 gives the mathematical definition of factor analysis in terms of the data matrix (X), the analysis itself can, as in regression analyses, proceed mathematically from the Pearson correlations among the variables. Factor analysis can be presented as an analysis of correlations without reference to actual scores, but that can be misleading. Some techniques that proceed from the correlation matrix (e.g., cluster analysis) have no direct mathematical relationship to the observed variables. Factor analysis does; it is an analysis of the observed data using correlations only as a convenient intermediate step. (Note that phi, Spearman rank, and point-biserial correlations are all special cases of the Pearson correlation coefficient and so are appropriate for factor analysis. Although other coefficients, such as biserial correlations, have been tried, they do not proceed directly from Equation 6.1 and can produce matrices that cannot be factored.)

Factor analysis could proceed from covariances instead of correlations. If covariances are used, then the variable with the largest variance is given more weight in the solution. For example, if income were measured in dollars per year and education measured in number of years spent in schooling, the former's variance would, being in the tens of thousands, influence the results much more than would the latter, whose variance would be less than 10. With social science data in which the variances are arbitrary, weighting the solution towards variables with higher variances is seldom useful. However, do note that correlations are affected by restriction of range. When the range is less than is normally found with a variable, the correlations are lower. When such restriction does occur, the factor loadings will be lower than when the range is larger. In such a situation, it is appropriate to either correct the correlations for the restriction of range or use covariances. Factoring covariances produces factor weights that are the same despite restrictions of range. However, they may, in addition to the inconvenient weighting, be more difficult to interpret because they are not in the range of -1 to 1 as are correlations. The discussion here assumes that correlations are being factored unless stated otherwise.

Table 6.1 gives a simple example of six variables (Gorsuch, 1983). The left part of the table gives the observed correlation matrix, and the second part gives the factors' correlations with

TABLE 6.1 Factor Analysis of 6 Variables

	r with Variables						*r* with Factors		
Variable	1.	2.	3.	4.	5.	6.	I.	II.	h^2
1. Information	—						.76	−.09	.59
2. Verbal ability	.67	—					.81	−.07	.66
3. Verbal analogies	.43	.49	—				.58	−.07	.34
4. Ego strength	.11	.12	.03	—			.06	−.67	.45
5. Guilt proneness	−.07	−.05	−.14	−.41	—		−.05	.59	.35
6. Tension	−.17	−.14	−.10	−.48	.40	—	−.12	.66	.45

Note. Correlation between factors = −.14.

the variables. They show that the first three variables form one factor and the second three form another. The reason the communalities are small is because these are all brief forms with low to moderate reliabilities in this sample.

The results of a factor analysis includes the degree to which each factor relates to each variable. When a factor relates to a variable, the common usage is to say that the factor *loads* the variable. Loading refers to the relationship of a factor to a variable in general but not to one particular numeric values. It is appropriate to use the term *loading* when one wishes to refer to whether the factor contributes to a variable. However, whenever a number is referred to, the type of factor loading must be reported. Thus it is appropriate to ask *Does factor A load variable 3?* and appropriate to respond *Yes, it correlates .58 with the variable.* There are three types of factor loadings. First are the weights for each factor's *z* scores to estimate the variable *z* scores. Second are the correlations of each factor with each variable. The last, and least used, is the partial correlation of each factor with each variable with the other factors partialled out. (These are discussed more in this chapter's section on correlated factor solutions.)

There is an assumption in least squares analyses of Equation 6.1, including factor analysis. Use of the model assumes that each equation applies equally to each person. It is difficult for these analyses to work well if the *X* is a function of Factors A and B for half the sample but a function of Factors C and D for the other half. Such may occur, for example, when there are multiple ways in which the variable can be changed. Consider a hypothetical situation in which children in poor communities only receive high exam scores if they are innately bright (because poor communities, we shall assume, cannot contribute much to their scores). Then those in rich communities would receive high exam scores less related to innate brightness because of the resources that led to a strong learning environment. Because different influences are at work in different parts of the sample, the factor analysis will be an averaged one and not represent either community well.

In factor analysis, the desire is to find a limited number of factors that will best reproduce the observed scores. These factors, when weighted, will then reproduce the observed scores in the original sample and, in new samples, will estimate what the observed scores would be if measured. Of course, the reverse may also be of interest: using the observed scores to measure the factor. But in the latter case, the factor is measured not to estimate the observed scores, but rather to generalize to other variables that also are correlated with the factor. These two approaches are seen in the examples. The boxes are analyzed to identify the factors: length, height, and width. Knowing the factors, we can in the future just measure length, height, and width directly and compute other variables such as a diagonal. The reverse is of interest in intelligence testing; scales such as Similarities and Vocabulary are used to measure verbal capability. Psychologists then examine, for example, a person's college grades in courses demanding high verbal capability to see whether they are as expected, given the person's verbal ability.

Note that in factor analysis, only the observed scores, the *X*s in Equation 6.1, are known; the factor scores (*A*, *B*, etc.), the weights (the *w*s), and the uniquenesses (*u*s) are unknown. With one known and three unknowns, it is mathematically impossible to solve for them without further restrictions. The restrictions adopted to allow solving for both factors and weights are a function of the factor model.

FACTOR ANALYTIC MODELS AND THEIR ANALYSES

To solve Equation 6.1 for both the factors and the weights, restrictions must be made. The restrictions can be minimal or extensive. The former—minimal restrictions—includes the class of models known as exploratory factor analysis (EFA). Mathematical principles are selected for the restrictions but there are no restrictions that take into account any theory that the investigator might have. The results are based solely on the observed data. The latter—extensive restrictions—includes the models known as confirmatory factor analysis (CFA). Based on theory or past research, a set of weights is proposed and tested as to whether the weights adequately reproduce the observed variables. Note that restrictions are not necessarily a dichotomy between minimal and extensive. Some forms of EFA are more restricted than others and some forms of CFA

are less restricted than others. These variations arise out of what the investigator is willing to or needs to specify.

Component Analysis

Component analysis (CA) restricts Equation 6.1 by dropping the uniqueness term, u. Thus the interest is in factors (also called components when using CA) that reproduce all of each and every variable, and so have expected communalities of 1.0. Of course, CA users would never argue their variables have reliabilities of 1.0 and so the actual maximum communality is generally much lower than 1.0. And CA users know the variables will not have multiple correlations of almost 1.0 with the other variables (needed for the factors to have a multiple correlation of 1.0 with each variable). Therefore no variable can, except by capitalization on chance, actually have a communality of 1.0. But proponents feel CA gives, with solid variables that correlate well, a reasonable approximation, with negligible distortion from the ignored unreliability and ignored multiple correlations less than 1.0.

Derivations easily show that the first step in all exploratory factor analyses is to compute the correlations among the observed variables. It is important to note that technically it is a covariance matrix among Z scores that is being factored. The main diagonal contains the variances—which are 1.0 by the definition of Z scores. The off-diagonal elements are technically the covariances among the Z scores which, because Z scores have variances of 1.0, are also the correlations among the variables. Procedures mentioned below are then applied to the correlation matrix to extract the components.

To extract factors from the data matrix, more restrictions need to be made than just assuming the us are zero. The restrictions are mathematical and use one of two procedures. The first, principal components, has the restriction that the first factor is the largest possible one, the second is the largest one after the first has been extracted, and so forth for all the factors. The second, maximum likelihood, adds the restriction that each should have the maximum likelihood of that found in the population. The latter is more difficult to compute, but both are quite similar—and both become more similar as the N increases. It would be surprising if there were any interpretable difference between these two procedures with a reasonable N.

The factors as extracted are seldom directly interpretable. Hence the factors are rotated (a term which comes from a geometric development of factor analysis; see Gorsuch, 1983, particularly chapter 4)—that is, are transformed to meet some criterion while keeping the same communalities. The usual criterion for rotation is simple structure, which can be briefly defined as the maximum number of variables loading only one factor with a side condition that these loadings be spread among as many factors as possible. Table 6.1 shows excellent simple structure. Each variable is loaded by only one factor and each factor loads a distinct set of variables. Because rotation applies to all EFA methods but has correlated and uncorrelated models in terms of how the factors are restricted, it is discussed further in the section of this chapter entitled "Restricting to Uncorrelated Factors" after the other EFA methods are noted.

CA is more parsimonious than are other models based on Equation 6.1 in that the equations are simpler when the unique term is dropped from Equation 6.1. One of the effects is that factor scores can be directly calculated (which, as noted below, is not true for the other major exploratory model, common factor analysis). These factors are linear combinations of the observed variables that can serve as summaries of the function represented by the factor. Such factors appeal to those who wish to stay close to the data and who philosophically hold that all constructs are just convenient summaries of data. (This is a discussion to which we return later.)

CA has been considered to be only an EFA procedure, with no CFA version. That is true within the narrower definition of factor analysis generally employed. But in terms of the model of Equation 6.1 and the logic of CA, a confirmatory components analysis is technically possible. The problem is that no significance tests are possible because the CA model has no place for errors.

Common Factor Analysis

Common factor (CFA) models use Equation 6.1, including the uniqueness term. Each uniqueness is the sum of several types of variance not in the factor analysis. These include random error (from unreliability and sampling error) and residual error in the sense that part of the variable is unrelated to the factors. The term *uniqueness* is used for all error because the random error, sampling error, and that which cannot be estimated from the factors can be considered unique to each variable. In CFA models, the focus is on the commonly shared variance of the variables and factors, hence the name common factor analysis.

Having the uniquenesses in the equations requires assumptions to restrict the analysis sufficiently for there to be a solution. These assumptions parallel those of residual-error-uniqueness in regression analysis. The uniquenesses are assumed to be both

- Uncorrelated with each other.
- Uncorrelated with the common factors.

Because nontrivial uniqueness may exist for each variable, the variance associated with the factors is reduced for each variable. The variables' Z scores have an original

variance of 1.0, but the part of each variable's Z scores that can be accounted for by the common factors is 1.0 minus u^2, and so will be less than 1.0. The importance of this for CFA is that the correlation matrix of the observed scores needs to be altered to take this into account. This is done by estimating the expected communality of each variable (because that is the squared multiple correlation of the factors with that variable and so is the variance of the reproduced variable) and replacing the 1.0 in the main diagonal of the correlation matrix with the communality. This is appropriate because the matrix is technically a covariance matrix, with the main diagonal elements being the variances of the variables.

Common factor analysis generally attracts those who wish to acknowledge the fact that all psychological variables have error and who prefer a model that is consistent with other methods of analysis, such as regression analysis and structural equations modeling. Factor scores, they originally felt, were not an issue because the factor score estimates correlate so high with the factors that the problem of factor scores' being only close approximations is minor; now proponents of common factor analysis suggest that factor scores are seldom needed because extension analysis can be used instead, and so the factor score issue is a moot question. (We return to the issue of CA vs. CFA later in this chapter.)

Common factor analysis has both an exploratory and a confirmatory model. An exploratory common factor analysis (ECFA) is one in which the restrictions are minimal both in number and in regard to the investigator's theories. It is an inductive analysis, with the results coming from the data as undisturbed by the investigator's thinking as possible. The advantage of not specifying an expectation is that the analysis is a multitailed test of any theory or expectation the investigator might have. If the investigator's expectations are found by ECFA, then they would certainly be found by a confirmatory analysis. However, due to the lack of restrictions and the complexities of the analyses, significance tests are not available for ECFA, so large Ns are to be used to reduce the need for significance tests.

Communalities could be calculated exactly if the factors were known and vice versa: The factors could be calculated exactly if the communalities were known. To cut this Gordian knot, the communality can be estimated and then the factors extracted. The observed communalities should differ only slightly from the estimated communalities.

Communality estimation is readily done by several methods. The following are four:

- *SMC:* Use the squared multiple correlation (SMC) of all other variables with that variable. This generally works well and is independent of the number of factors.
- *Pseudoiteration:* Use anything as the initial estimate, solve for the number of factors (see the following discussion for how to estimate the number of factors), and calculate the communalities from these factors. Then use the observed communalities as new estimates of the communalities, extract factors again, and calculate the communalities from these factors. Continue the process until little change is noted from one pass to the next or a maximum number of passes has made. Note that this is *not* true iteration. True iteration occurs when it has been proven both that the iterated values necessarily converge and that they necessarily converge to the right values. But neither necessarily happens with pseudoiteration. Gorsuch (1974, 1983) has noted a case in which the process would not converge, so the requirement for true iteration that the values converge is not met. The condition that they converge to the right values is not met because they sometimes converge to an impossibly large value. For example, in practice, communalities computed by this process often exceed 1.0. (Values greater than 1.0 are referred to as Heywood cases after the author of the first published discussion of the situation. Actually, those using the criterion of 1.0 to conclude the estimates are incorrect are optimists; the actual upper limit for communalities are the reliabilities of the variables, which are almost always less than 1.0. Thus, more violations of the upper limit occur than just the Heywood cases.) The fact that the process need not converge to values that are possible means this process is not an iterative process in the mathematical sense. In mathematics a procedure is iterative if and only if it is found to converge on the population value. Therefore the so-called iteration for communalities is only pseudoiteration. Why is pseudoiteration widely used? I suspect that there are two reasons. First, mathematical iteration is an excellent procedure, so iteration was certainly worth a try even though there is no mathematical proof it meets mathematical criteria for iteration. Second, when starting from 1.0 as the initial communality estimate, we see that the first few pseudoiterations obviously lower the communality estimates from the too-high value of 1.0 to a more reasonable estimate.
- *SMCs with two to three iterations:* This procedure starts with the SMC noted previously. Then the solution is iterated two or three times and stopped. Although it is still a pseudoiteration, it has never in my usage produced an estimate over 1.0. Snook and Gorsuch (1989) found the resulting communalities to not differ significantly from the communalities designed into the study. This is a good procedure.

- *Minres analysis:* This procedure minimizes the off-diagonal elements while using no communality estimates. Communalities result from the analysis. It is an excellent procedure if exact communalities are desired.

Some of the concern with communality estimates has been found to be an overconcern. Any reasonable estimate (plus several other similar ones, including special adaptations of b. in CFA) produces a final solution that is indistinguishable from the others. This is probably the reason that Minres is seldom used.

Note that the number of elements of the main diagonal of the correlation matrix—which are replaced with the communality estimates—increases linearly with the number of variables, while the number of nondiagonal elements increases much faster. For example, with six variables the communality estimates form 29% of the values being analyzed. With 30 variables, the communalities form only 7%. With 60 variables, the percentage is down to 4%. The impact of the communality estimates becomes increasingly unimportant as the number of variables increases.

In addition to the number of variables, a second parameter that is important in evaluating the importance of the communality estimates is how high the communalities are. The higher they are, the narrower the range of estimates for the communalities. With higher communalities, it is less likely that using a different communality estimation procedure would result in an interpretable difference.

Table 6.2 contains communalities for the box, WAIS, and psychological variable examples. They were computed from three initial estimates, 1.0, SMC, and SMC plus two iterations. The resulting communalities from the factors based on each estimation procedure are given. (The 1.0 column contains the actual communities from component analysis even though they were assumed to be 1.0.)

For the psychological variables—where the communality estimates are low to moderate and form 29% of the coefficients being analyzed—using 1.0 as the initial communality estimate makes a difference, but there is little difference between the other two initial estimates. In both the box and the WAIS examples, the communalities are high, so the estimates give quite similar results. Table 6.2 contains the factor loadings for the SMR plus the two-iterations solution for the six psychological variables data set.

Any of the parameters of Equation 6.1 can be zero. Now note what happens if the variables have high multiple correlations with the other variables. As the multiple correlations increase, the uniquenesses, *u*s, approach zero. If they were zero, then the *u*s would drop out and it would be a CA. Hence, CA is a special case of ECFA. An unrestricted ECFA will give CA if the variables have high multiple correlations with each other. (It is for this reason that CA and ECFA are part of the same statistical model even though it they may be used for different purposes.)

TABLE 6.2 EFA Communalities Using Different Initial Values

Estimation:	1.0	SMR	SMR + 2 Iterations
	Psychological Variables (2 Factors)		
1.	.73	.55	.59
2.	.77	.61	.66
3.	.56	.34	.34
4.	.65	.41	.45
5.	.57	.33	.35
6.	.64	.41	.45
	Boxes (3 Factors)		
1.	.95	.93	.91
2.	.96	.93	.93
3.	.93	.97	.98
4.	.96	.98	.98
5.	.97	.99	.99
6.	.98	.98	.99
7.	.91	.90	.88
8.	.87	.84	.82
9.	.98	.97	.97
10.	.90	.73	.68
	WAIS-III (4 Factors)		
1.	.65	.59	.60
2.	.70	.51	.54
3.	.76	.42	.46
4.	.81	.44	.49
5.	.79	.65	.66
6.	.65	.52	.55
7.	.69	.40	.43
8.	.63	.34	.35
9.	.77	.68	.69
10.	.76	.51	.56
11.	.84	.74	.77

As is the case with CA, ECFA proceeds by extracting factors by principal or maximal likelihood methods. The restrictions are then changed in the rotation of the factors (mentioned in the discussion of CA and discussed further later in this chapter). For example, the rotation reduces the number of factors loading each variable so that the relationships will be simpler than if most factors loaded most variables.

Confirmatory Common Factor Analysis

Confirmatory common factor analysis (CCFA) has been developed and used within the common factor model. It proceeds directly from equation 6.1 and includes the uniquenesses. But unlike ECFA, which uses mathematical restrictions to gain a solution, confirmatory methods use theory to develop appropriate restrictions.

The restrictions can be placed on any or all of the following of Equation 6.1:

- The number of factors.
- The weights of a factor to reproduce a variable.
- The uniqueness for each variable.
- The means and standard deviations of the factor scores. These are generally set to either Z scores (mean = 0, SD = 1) or the mean and SD of a particular variable.

It is possible also to place restrictions in addition to the elements of equation 6.1. The prime such restrictions are on the following:

- The correlations (or covariances, if covariances are being analyzed) among the factors.
- The correlations (or covariances) among the uniquenesses. These are generally restricted to 0, but they can be placed at other values. If non-zero, they can represent correlated method or errors.

The restrictions vary in the values that can be used. The more useful variations are to restrict a parameter to 1 or 0. When the means and standard deviations of the factors are set to 0 and 1, respectively, the factors are then Z scores. The correlations among the factors are set to 0 to restrict the factors to being uncorrelated.

The weights can be restricted in multiple ways. Here are the usual weight restrictions:

- The most widely used weight restriction is to set some weights to 0. This means that the variable is defined without regard to that factor.
- A predefined weight may be used; this is useful in evaluating whether the weights from another study are cross-validated in the current study.
- Several weights can be restricted to being the same value, with the value not predefined; for example, this is used if one has two parallel forms of the same measure.

If the weight is unrestricted, then the factor extracted is expected to have a nonzero weight on that variable, and the investigator wishes to know if that is so. The number of restrictions must be sufficient to identify a unique solution. Identification can be a problem in that no one has yet developed a formula to say when a unique solution is identified. It has been impossible to give a specific answer because the value depends on not just the number of restrictions but also their location. However, a correlation-based CCFA is generally sufficiently restricted if each variable is only allowed to be loaded by one factor and each factor has at least three

TABLE 6.3 Confirmatory Common Factor Analysis of 6 Psychological Variable Problem

	Hypothesized Weights		ECFA Principal Factor Weights		CCFA Weights (maximum likelihood)	
1.	?	0	.77	0	.77*	0
2.	?	0	.87	0	.86*	0
3.	?	0	.57	0	.57*	0
4.	0	?	0	−.70	0	−.70*
5.	0	?	0	.58	0	.58*
6.	0	?	0	.69	0	.70*
	r = ?		r = −.12		r = −.21*	

Note. * $p < .05$. ? means the value is left free to vary.

such variables. Usually the computer program reports any problems occurring that could be caused by insufficient restricting, referred to as *underidentification.*

For a CFA example, consider the six psychological variable example. From general psychological knowledge, we would expect that any factor of the verbal ability measures would not load the psychological distress variables, and vice versa. Hence, the hypothesized pattern would have six values set to zero. The other three values for each factor would be allowed to vary (i.e., would be set by the program). The correlation between the factors is unrestricted (see Table 6.3).

Consider just the first factor in Table 6.3. What the restrictions in the hypothesized weights say is that the last three variables are not to be considered in the solution of that factor. But it does not say how the weights for the first three variables are to be found. What is needed is the factor that best reproduces the scores of these three variables. Note that this is the same question asked in ECFA, and the same restriction is used so that a solution can be found: *principal factoring* (maximizing the variance that is reproduced) or *maximum likelihood factoring* (maximizing the variance with the further restriction of maximizing the generalization to the population). To illustrate this connection with ECFA, one principal factor was extracted from the first three variables; then, separately, one factor was extracted from the last three using an ECFA program (communalities were started at reliabilities and then iterated nine times). That is the second part of Table 6.3. It gives the weights for each of the factors to reproduce each of the variables. Using extension analysis (discussed later in this chapter), the correlation between these two so-called exploratory factors was found to be −.12.

And what if a real CCFA is computed from these data? Using the original maximum likelihood program for CCFA gives the final two columns of Table 6.3. The very slight differences may be a function of the differences between principal and maximum likelihood factors or the number of iterations for communalities. (It does illustrate how few

differences there can be between principal and maximum likelihood factors.)

There is a warning in the use of CCFA: Changing the parameters of the model after looking at the data may well lead to a nonreplicable solution. The model needs to be set before the analysis begins. If more than one model needs to be tested, then all models need to be completely specified in advance.

If a hypothesized CCFA model gives a less-than-desired fit to the data, investigators occasionally make some adjustments to produce a better fitting model. This is a dangerous practice because it capitalizes on chance. The literature suggests such changes often lead the model away from the population model, not towards it. None of the significance tests nor the goodness-of-fit measures take this capitalization into account. If any changes are made to improve the fit, the report needs to explicitly state the original model, give the basis for all changes, and warn that some capitalization on chance will have occurred. It is recommended that a cross-validation sample be used to test any model containing data-based changes.

What is the advantage of a real CCFA over just extracting factors from subsets of the variables? The answer is *significance tests.* In Table 6.3, the CCFA found all the loadings to be statistically significant. These significance tests are possible because the solution is sufficiently restricted to be mathematically tractable.

Restricting to Uncorrelated Model Factors

The previous discussion of component and common factor models fits the general case in which there are no restrictions on the correlations among the factors. This is appropriate in most cases because either the variables are all drawn from the same domain, or how the domains relate is of interest. But allowing for correlations among the factors adds some complexity.

The simplicity introduced by uncorrelated factors is the same as with uncorrelated predictors in multiple regression. Multiple regression analysis simplifies if the predictors are uncorrelated with each other. With uncorrelated predictors,

- The correlation of the independent variable with the dependent variable is also its Z score weight, and its correlation when all the other predictors are partialed out (the partial correlation).
- There is no overlapping variance among the independent variables, so the correlation is unchanged if one of the other independent variables is partialed out or is not in the equation.
- The multiple correlation is the square root of the sum of the squared correlations of the independent variables with the dependent variable.

In factor analysis, the factors are the predictors or independent variables, the observed variables are the dependent variables, and the communalities are the squared multiple correlations of the factors with the observed variables. Thus, with uncorrelated factors,

- The correlation of the factor with an observed variable is also its Z score weight, and its correlation when all the other factors are partialled out (the partial correlation).
- There is no overlapping variance among the factors, so the correlation is unchanged if one of the other factors is partialled out or is not in the equation. However, because the uncorrelated restriction is applied to this specific set of factors, dropping a factor from the solution can change the weights.
- The communality is the square root of the sum of the squared correlations of the factors with the variable.

Because the correlation is equal to the weight and is equal to the partial correlation, there is only one interpretation for the term *loading* when the factors are uncorrelated. With correlated predictors or factors, the three conditions previously noted do not hold. Instead the beta weight (in regression analysis; factor weight in factor analysis) differs from the correlation, and those differ from the partial correlation (when the other predictors/factors are held constant). The multiple correlation/communality is computed by a more complex formula that takes the correlations among the variables/factors into account.

In factor analysis with correlated factors, each type of loading is put into a separate matrix. These have been named

- The factor pattern that contains the beta weights given to the factor Z scores to reproduce the variable Z scores.
- The factor structure that contains the correlations of the factors with the variables.
- The reference vector structure that contains the correlations of each factor with the variables with all other factors partialled out.

The factor pattern is generally considered to be the one to interpret, but the other matrices can be of interest also. Often the reference vector structure is clearer than that of the others because the correlations of factors with variables solely due to how the factors intercorrelate have been removed.

Because uncorrelated factors are easier to work with, why not restrict all factor solutions to being uncorrelated? The answer is that it may lead to a misleading representation of the data. For example, ability scales are all generally correlated together. This is true of the WAIS-III data; the lowest correlation is .22 (Digit Span with Digit Symbol) but correlations in the .50s and .60s are common. This is true not only among the scales, but also among the IQ and Index scores. Restricting to uncorrelated factors fails to inform us that the abilities are highly related.

Solutions restricted to uncorrelated factors are also referred to as orthogonal, a term from the geometric representation of factors. In the same manner, unrestricted solutions are also referred to as oblique. However, that term can be misleading. It implies that the solution is restricted to having correlated factors, which is not the case. Unrestricted rotation is just that: unrestricted. Factors can and often are uncorrelated when unrestricted factor rotation is used.

Many procedures exist for rotating factors, but the decision usually is just whether the factors will, on an a priori basis, be restricted to being orthogonal or will be unrestricted. If restricted, the program of everyone's choice is Varimax. For unrestricted rotation, there are several options, with most giving reasonable solutions. Some such as Oblimax have a parameter to set that influences the degree to which the solution is forced towards orthogonality. The most elegant unrestricted rotation is to start with Varimax, and then use Promax to provide an unrestricted version of the Varimax solution. Like other unrestricted solutions, there is a parameter to be set, referred to as k. Part of Promax's advantage is that the value of k is no longer a choice to be made because it makes little difference . It can always set to 4. With this setting, uncorrelated factors will result if appropriate, because orthogonal rotation is a special case of unrestricted rotation.

Note that Promax may produce factors with correlations so trivial that they can be treated as uncorrelated factors, as in Table 6.1 in which the correlation was a trivial −.14. Milliron (1996) found in a simulation study that Promax was good not only for correlated factors, but also replicated the known factor pattern better than Varimax did for factors uncorrelated in the population. In the samples, Varimax had to slightly distort the loadings to keep the factors correlating exactly zero, whereas Promax allowed for chance correlations among the factors.

Occasionally there are unexpected results with Varimax. Not only is an obvious general factor completely missed, but also the zero correlations among the factors can disappear at the next calculation. Several studies have used Varimax and then estimated factor scores. The factor scores were obviously correlated, indicating that the restriction could not be applied through all the calculations because the restricted rotation fit the data so poorly. Other studies have used the orthogonal factors of a prior study in a new sample, only to find the factors correlating .6 to .8. Highly correlated data will not be denied. It is best to be forewarned about this situation by leaving the rotation unrestricted.

If the factors are correlated, then those correlations can be factored (just as the original variable correlations were factored). The factors from the variables themselves are called the primary factors, whereas those extracted from the primary factors are called secondary factors; third-order factors would be factors from the second-order factors, and so forth. All factors after the primary factors are referred to as *higher-order factors.* Conceptually, the primary factors are more specific than are the secondary factors and so should predict more specific variables better than do the secondary factors. With more general variables, the secondary factors should predict better. Using the results of a higher-order factor analysis and the desired dependent variables, it is possible to show (Gorsuch, 1984) and even test (Mershon & Gorsuch, 1988) when the primary or second-order factors are more useful.

An example of higher order factoring is the WAIS-III. The primary factors are in Table 6.4. The four primary factors were correlated, and a general second-order factor was extracted. This factor, the last column of Table 6.4, represents the classical g, or general ability factor (IQ). The correlations of the individual scales with g were computed by extension analysis (discussed later in this chapter). It is g that has a long history of relating to many areas of achievement.

TABLE 6.4 Higher-Order Analysis of the WAIS-III (Canadian) First-Order Factors and Correlations of the Primary Factors

Variables	1. Verbal Comprehension	2. Processing Speed	3. Working Memory	4. Perceptual Organization	*g*
Arithmetic	.25	.02	.38	.26	.69
Block design	−.08	.11	.03	.70	.63
Digit span	−.01	−.02	.70	−.01	.47
Digit symbol	.05	.69	.00	−.03	.46
Information	.75	−.05	.09	.04	.63
Matrix reasoning	.07	−.03	.05	.68	.64
Letter number cancellation	−.03	.07	.61	.03	.48
Picture completion	.10	.01	.01	.51	.41
Similarities	.69	.05	−.08	.22	.46
Symbol search	−.01	.68	.04	.09	.54
Vocabulary	.85	.06	.03	−.03	.66
Correlations of the primary factors					
1.	1.00	.44	.51	.67	.73
2.	.44	1.00	.50	.56	.65
3.	.51	.50	1.00	.60	.71
4.	.67	.56	.60	1.00	.85

Note. The first-order and second-order factors used SMRs plus 2 iterations as communality estimates for the ECFA using principal factors extraction and Promax rotation. The correlations of g (i.e., the general factor) with the scales was by extension analysis (Gorsuch, 1997).

If one suspects that there is a general factor and CA or ECFA is used, that general factor will usually be found if and only if a higher-order analysis is computed from unrestricted rotation.

Item analysis is probably the most common situation in which a rotation restricted to orthogonality is misleading. The author of a scale includes items that each measure the underlying characteristic; then a total score is computed by adding the items together. So the author is assuming that there is a general factor—that is, one that loads all of the items. What happens when the scale is factored? Because factor analysis is a sensitive tool, it will take into account the almost universal fact that some items will correlate more highly with each other than with the rest of the items. There are generally several subsets of items that correlate slightly higher among themselves than with the other items because they have the same distributions or use similar words. Then several factors will be found. These factors may, for example, be one for the easy items, one for the medium-difficulty items, and one for the hard items. None of these factors will be a general factor because, as in Table 6.4, the general factor is found in the correlations among the factors. Varimax, however, never allows such correlations to occur. The decision to restrict item analysis rotation to orthogonality is a decision with major implications. It is far better to use Promax, an unrestricted rotation, and see whether a general factor happens to occur among the factors.

An instructive example can be drawn from the factor analyses of the Beck Depression Inventory (BDI). Chan (Gorsuch & Chan, 1991) ran analyses in Chinese and U.S. samples, and computed the relationships of previous U.S. and Canadian factor analyses to her factors. The table clearly showed that (a) primary factors did not replicate, whether within or across countries; (b) all primary factors correlated highly; and (c) the second-order depression factor replicated both within and across countries. That general factor is the same as the total score. The prior studies missed this fact because they only provided first-order analyses, and the erroneous conclusion from those would have been that there were no replicable factors. Chan showed the correct conclusion to be that there is one factor in the BDI, just as the author designed it.

MAJOR UNRESOLVED ISSUES

In the previous discussion, suggestions have been made for computing a factor analysis using reasonable and generally accepted solutions. These include using Promax unrestricted rotation. Also widely acceptable are squared multiple correlations with two iterations for communality estimation (although pseudoiteration is most widely used, and is alright until it gives communalities higher than the observed reliabilities). But some major issues are currently being debated with little common agreement on their resolution, although there is evidence to evaluate the usefulness of different methods.

Two methods are used to evaluate the usefulness of a factor analytic technique. These are simulation studies and plasmodes (Cattell, 1978). Simulation studies start with a population factor pattern and factor correlations as givens (they are selected by the investigator to be sensitive to the parameter being investigated). The pattern and correlations may be systematically varied. Then hundreds to thousands of samples are derived using the population parameters, but allowing chance variations due to sampling. These multiple samples are analyzed, and the conditions under which the selected parameters are best recovered are noted.

Plasmodes are data sets in which it can be reasonably assumed that we know what the results should be. The examples used in this chapter fit that category. The history of psychology suggests that verbal ability and emotional distress are separate factors (the six psychological variables), and who would question the need for factors of length, height, and width to underlie boxes? The WAIS family of ability measures, of which the WAIS-III Canadian data set is one example, has a long history of factor analysis; the four-factor solution presented previously was replicated with multiple samples across both the WISC and WAIS. Which of several competing factor analytic techniques most ably find the expected results?

Although it is easy to vary parameters in simulation studies, there is always the question of generalization to the type of data commonly analyzed. And although plasmodes are data like those commonly analyzed, it is difficult to systematically vary parameters. Hence, our discussion of the problem areas relies heavily on both simulation studies and the plasmodes already presented as examples in this chapter.

What is the final arbitrator of factor analytic methodology? The ultimate arbitrator in science is well established: replication. Any procedure that produces replicable results is worthy of consideration. If several procedures lead to replicable results, then the choice is based on fit to the investigator's theory and situation. If there is still a choice, then parsimony and elegance are the deciding factors.

Component Versus Common Factor Models for Exploratory Factor Analysis

Both CA and CFA are used for EFA. Although the existence of two models is not surprising, the level of debate has been extensive. For detailed discussions of the pros and cons of

these two models, see the special issue of *Multivariate Behavioral Research,* 1990, Volume 25, Issue 1 (also see 1996, Volume 31, Issue 4 for discussion of indeterminacy per se).

In understanding this debate, it is important to note that all procedures for CA and ECFA are the same except for one: CA starts with 1.0 in the main diagonal of the correlation matrix and CFA starts with a communality estimate (thus taking into account the existence of variance unique to the single variable). This is the only mathematical difference between the two. Everything else is the same (which is why they are both special cases of the general factor analytic model).

The Case for Common Factor Analysis

The rationale for CFA comes from Equation 6.1 and assumptions about data. Including the uniqueness term in the equation makes it a CFA. The uniqueness term includes all of the variable's variance not associated with the factors, part of which is random error. So, the CFA rationale goes, CFA should be used whenever at least some reliabilities are less than 1.0—that is, whenever some variables contain any random error. Of course, this argument runs, who can show, or *assume,* that all their variables are without random error? Where is the evidence for such variables in the social sciences? And if we know the variables have error, is it not rational to build that into our mathematical models?

Dropping the uniqueness term also means that the factors and only the factors underlie the scores for each variable. Hence in the population, the communality is to be 1.0. This is the justification for using 1.0 in the main diagonal of the correlation matrix. This means that the multiple correlation of the factors with each of the variables is also 1.0. Unfortunately, the derivative is that the variables, being sets of linear combinations of a fewer number of factors, will form a non-Gramian correlation matrix. Such a matrix has an infinite number of solutions and so cannot be factored at all. Therefore, CA is a self-contradictory model. (The only reason that CA works is that the model is wrong for the data—no two of the variables being analyzed have a multiple correlation of 1.0 with the same factors, so none truly fit the model.)

Although component advocates raise the problem of estimating communalities and factor scores, such estimates are consistent and easily made. The variations on factor scores are variations among scores that generally correlate .9 or better in simulation and plasmode studies. This is much better than in other areas. For example, major ability tests often correlate .7 to .8, yet are seen as interchangeable. Also the correlation between CA factor scores from one study to the next is much less than 1.0 and is probably no greater then that from one CFA to another, so where is the added precision

TABLE 6.5 Component Analysis of 10 Variables: Promax Factor Pattern

	Factors		
Variable	1	2	3
1. Length squared	.34	−.72	.09
2. Height squared	.57	−.26	−.14
3. Width squared	.13	.51	.16
4. Length + width	.49	−.12	.26
5. Length + height	.07	.61	−.06
6. Width + height	.40	.18	.14
Inner diagonals			
7. Longest	−.24	−.12	.44
8. Shortest	.59	.08	−.07
9. Space	.10	−.04	−.48
10. Edge thickness	.26	−.02	.80

Note. $N = 100$.

from CA? And with extension analysis (discussed later in this chapter), there is no need to compute factor scores because the correlations of variables not in the factor analysis with the factors can be mathematically computed.

The ECFA versus CA is a real question because the results vary dramatically in a few special situations. Table 6.5 presents the results of a CA. Factor 1 has two to four good markers, Factor 2 has two excellent and one good loading, and Factor 3 has one excellent and two moderate loadings. The loadings are clear and both the author and the reader would interpret them.

Unfortunately the matrix from which Table 6.5 was computed has not a single significant correlation. Each and every multiple correlation of one variable with the rest is, when shrunken for capitalization on chance, zero. The high loadings come from the assumption that all the variance of each variable is to be reproduced by the factors. Although this may be an unusual case, ECFA is better at protecting the discipline from such data than is CA.

There is also the principle of parsimony and elegance. That mathematical model is more elegant when it accounts for a wider range of situations. Equation 6.1 with the uniqueness term is using the same model as regression analysis, CCFA, structural equations modeling, and all other least squares techniques. To introduce a new model is to reduce parsimony and elegance among our statistical models.

The Case for Component Analysis

CA is more parsimonious because its equation is simpler. That makes it easier to teach and easier to program.

But the major arguments for CA go beyond having a simpler equation. One such rationale is a philosophical one. Factors are abstractions from data that we make for our convenience, not to be reified into realities. Factors are just

TABLE 6.6 Component Analysis of WAIS-III Canadian Data

Variables	Components			
	1	2	3	4
Arithmetic	.39	.04	.36	.21
Block design	−.09	.13	.02	.81
Digit span	.00	−.06	.91	−.04
Digit symbol	.06	.93	−.04	−.08
Information	.91	−.05	.04	−.03
Matrix reasoning	.14	−.03	.04	.71
Letter-number cancellation	−.06	.06	.83	.01
Picture completion	−.01	−.09	−.07	.86
Similarities	.80	.04	−.10	.16
Symbol search	−.04	.82	.03	.12
Vocabulary	.93	.04	−.01	−.04

TABLE 6.7 Correlations Between Component Analysis and Exploratory Common Factor Analysis Solutions

	CA Solution								
	Psychological Variables		Boxes			WAIS-III			
	1.	2.	1.	2.	3.	1.	2.	3.	4.
ECFA Solution									
Psychological Variables									
1.	.88	−.20							
2.	−.18	.82							
Boxes									
1.			.99	.67	.65				
2.			.60	.95	.62				
3.			.70	.74	.96				
WAIS-III									
1.						.93	.45	.54	.68
2.						.38	.81	.44	.49
3.						.42	.43	.81	.52
4.						.59	.51	.55	.85

Note. ECFA was with SMRs plus 2 iterations for communities. Correlations computed by extension analysis (Gorsuch, 1997).

that—convenient constructs that help our generation relate to the data consistencies we find in our discipline. And because they are *our* constructs, we choose to define them by the CA model.

Another rationale for CA is a set of pragmatics. One such pragmatic is that using CA instead of CFA seldom makes much difference. Many factor analyses are of 25 or more variables with, if the study is designed well, reasonably high communalities. In such cases, the results of CA and CFA lead to the same conclusions. Compare the CA in Table 6.6 against the CFA of Table 6.4. Is there really an interpretable difference? And in fact do not the high loadings stand out better from the low ones in the CA?

Other rationales for CA arise as much from classical limitations of CFA as from the CA model. A major limitation arises from the communality problem. Because we never know the communalities but only estimate them, there are a set of solutions that fit the data equally well. And iterating for communalities can produce Heywood cases.

As the communalities can only be estimated, the further mathematical conclusion is that there are an infinite number of factor scores that could be computed that would fulfill the ECFA model equally well for any given data set (a result of what is called the indeterminacy problem). With CA, the factor scores are a linear combination of the variables of which there is only one set.

The Ongoing Debate

While the existence and use of two models is not surprising, the level of debate is surprising. The results from both are, except in special cases, quite similar. Table 6.7 gives the correlations between the factors of CA and ECFA for the three examples. Particularly instructive is the psychological variables example. It has the fewest variables and the lowest communalities, which are the conditions under which the CA and CFA might be expected to differ. It seems that the replication of factors between CA and ECFA are good for the six psychological variables and excellent for the other two data sets. These are so high that we would be delighted to get them if testing for replication from one sample to another within either CA or CFA.

Personally, I had the good fortune both to study with a major exponent of CFA (Cattell, 1978) and to work with a major exponent of CA (Nunnally, 1967), both scholars I respect highly. The former was my mentor in graduate school; I was employed by the latter to calculate all the examples for his book and gave paragraph-by-paragraph feedback on it. (Nunnally returned the favor by providing paragraph-by-paragraph feedback on the first edition of my *Factor Analysis;* Gorsuch, 1974.) So I heard both arguments multiple times. And in following the dialogue for the past 30 years, the only major change seems to be that the heat of the debate has increased.

Professional debates are good, but the search is (should be?) for procedures that address the critiques of both sides. I proposed such in the *Multivariate Behavioral Research* special issue (Vol. 25(1); Gorsuch, 1990): image analysis. Image analysis is a special case of common factor analysis, which factors the part of the variable that correlates with the other variables. Thus, it is oriented toward the common factors (i.e., factors that load at least two variables). The part that does not relate to another variable is dropped from the model. Thus, image analysis includes all that the supporters of ECFA want. This should satisfy the proponents of ECFA. For the

proponents of CA, image analysis answers their critiques of ECFA because there is no communality problem and factor scores can be calculated, not estimated. Thus image analysis should satisfy both the common factor and component advocates. Except for Velicer and Jackson (1990), this suggestion was ignored.

Personally, I opt for CFA for two reasons. First, including the uniqueness term means that the same equation is used for factor analysis as is used for regression and SEM (structural equations modeling). Second, a procedure should be as failsafe as possible, which means that loadings based on random correlations (Table 6.5) should look low to reduce the chance of believing there are significant loadings when there are no significant correlations. The issues of estimating communalities and estimating factor scores are, with contemporary procedures, trivial issues; the results correlate so highly that these are not problems. I do find it interesting that CFA was the original mode of factor analysis. Little if any consideration of CA is found before 1960. Instead common factor analysis was assumed and that was the only model presented. Insomuch as component analysis appeared at this point, it was just a special case of common factor analysis.

In 1960 computers entered psychology, but they were simple and slow. In illustration, the 1960 computer was slower and had less memory than the first Apple personal computer. Hence all programs had to be kept simple—very simple. It was then Henry Kaiser at the University of Illinois introduced the simplest complete computer package, called "Little Jiffy." It was doable in those computers because it was CA and had no communality estimation procedure (pseudo-iterated communalities would have literally taken too long for students to run). In his later discussions of this, he indicated that it was an oversimplified model. In 1970 (Kaiser, 1970) he introduced "A Second Generation Little Jiffy" but then it was too late. The computer packages had already picked up the runable "Little Jiffy" and that is still often the default in major statistical packages. My personal opinion is that the rationales for CA developed as a post hoc explanation because so many used a computer package which had "Little Jiffy" as the default. BUT NOTE: the origin of any construct in science is not judged by its history but only by its merits.

An important point to me is that CA versus CFA is a minor point with a reasonable number of variables and reasonable communalities. They give the same conclusions regardless of the philosophical or theoretical model the investigator wishes to assume. Only with a limited number of variables is there a difference, and then the best solution seems to be CFA because CA can make insignificant correlations into loadings that appear major. Much more important are issues such as variable selection, sample of cases, the number of factors to extract, whether there is warrant to restrict the solution to uncorrelated factors, and whether to run confirmatory or exploratory analyses. Particularly important is underestimating the number of factors (see the next section) and any decision to restrict the rotation to uncorrelated factors.

Number of Factors Issue

In the proceeding discussions, the number of factors has been assumed. That was to enable the major points of the models to be presented. Unfortunately, there is no adequate way of determining the number of factors in either exploratory or confirmatory factor analysis. It is not for want of trying, for numerous proposals have been made and numerous simulations studies have been run (Velicer, Eaton, & Fava, 2000, summarizes the results of the simulation studies for CA and EFA). Generally, it is recommended that the user examine several of the following procedures in setting the number of factors.

The following tests are only a sample of the total available and include the most widespread and those with the best simulation results.

Eigenvalue/Characteristic Root Criteria. From a correlation matrix eigenvalues can be extracted (formerly the common name for eigenvalues was *characteristic roots,* which is why the criteria in this section use the term *roots* so often). These have many characteristics, with the important one (for the present purposes) being that they are the sum of squared correlations of the variables with a principal or maximum likelihood factor. Each of these factors accounts for the maximum amount of the variance of the correlation matrix. They are extracted in order of size. Hence, the set of roots for a problem gives the sizes of the extracted factors from the largest to the smallest. (Note: Rotated factors have no roots; the term and theory apply only to factors extracted from the correlation matrix with 1.0 in the main diagonal because the estimated communalities depend on the number of factors. All were originally developed for the CF model.)

The roots for each of our examples are in Table 6.8. They are ranked in order of size, and show the pattern typical of roots of correlation matrices.

To put the roots into perspective, consider what the roots would be if there were no factors at all. In that case, the correlation matrix would have the variable correlations (off-diagonal elements) all equal to zero while the diagonal elements would be 1.0. A legitimate solution would be with the first extracted factor loading the first variable 1.0, with all other loadings being zero. This root, the sum of the squared loadings, would be 1.0. The second factor would be the

TABLE 6.8 Roots for Example Problems

Extracted Factor	Psychological Variables	Boxes	WAIS-III
1	2.30	8.22	5.36
2	1.63	.78	1.06
3	.71	.39	.86
4	.53	.31	.80
5	.51	.18	.64
6	.32	.05	.60
7	—	.03	.43
8	—	.02	.40
9	—	.02	.35
10	—	.01	.26
11	—	—	.23

second variable, with a loading of 1.0 and a root of 1.0. The rest of the factors would follow the same pattern, and all roots would be 1.0.

Roots Greater Than 1.0

Because all roots would be 1.0 in a matrix with no factors, one suggestion is that any root greater than 1.0 will reflect a value greater than zero in the off-diagonal elements and so will be variance that can be attributed to a common factor. In actuality, smaller roots may also reflect correlations, so technically roots greater than 1 is the minimum number of factors to extract, but common usage treats it as the number of factors to extract. This has been the most widely programmed, and so the most widely used, of all the criteria. Unfortunately, the simulation studies have found it to be the prime candidate for the worst criterion ever tried (Gorsuch, 1983; Velicer et al., 2000). In our examples, it is only correct with the psychological variables.

Parallel Analysis

The rationale of roots greater than 1 is for the population matrix, not for a sample matrix. All sample matrices will have random correlations that will produce roots greater than 1. Parallel analysis consists of doing parallel analyses of random data. They are parallel in that the same number of cases and variables are used as in the factor analytic study, but they consist of random data only. Fifty to 100 of these are run, and the roots are averaged to show what the roots would be if the data were only random. The roots always start over 1.0 and then drop fairly sharply. The larger the N, the flatter the slope of the roots.

Tables (Lauhenschlagen, Lance, & Flaherty, 1989) have been provided so that each person does not need to compute multiple analyses of random data. Equations can also be used (Velicer et al., 2000). In each of these cases, the parallelism is established by having the same number of variables and cases. It may be more appropriate to base the parallel analyses on matrices that also match the observed data in skew and kurtosis as well.

All roots from the factors of the study that are larger than the same numbered averaged random root are considered valid roots. For example, for the psychological problem with six variables and $N = 147$, the closest tabled values give the first parallel roots as 1.2, 1.1, and 1.0. The first observed root of Table 6.8 is larger than 1.2 and the second is larger than 1.1, but the third is less than 1.0. Therefore, parallel analysis indicates that two factors should be extracted because there are only two roots that exceed their randomly based equivalent. For the box problem, it gives one factor instead of three. The number of WAIS factors is also underestimated, giving two instead of four. It has serious problems with small but replicable factors.

Simulation studies have found parallel analysis to be a prime candidate for the best procedure for estimating the number of exploratory factors.

Scree Test

The scree test has a somewhat different logic for use of the roots. It is assumed that the variables cover a domain of interest and have at least moderately strong correlations. That means the factors of interest should be noticeably stronger than the factors of little interest, including random correlations. So when the roots are plotted in order of size, the factors of interest will appear first and be obviously larger than the trivial and error roots. The number of factors is that point at which the line formed by plotting the roots from largest to smallest stops dropping and levels out.

The name is from an analogy. *Scree* refers to the rubble at the bottom of a cliff. The cliff itself is identified because it drops sharply. The last part of the cliff that can be seen is where it disappears into the scree, which has a much more gradual slope. Note that the cliff is still seen at the top of the rubble; in the same way the number of factors includes the last factor associated with the drop.

Following the suggested use of the scree test gives three factors for the psychological variables and four for the boxes. That is one more than are assumed to exist in these two data sets. For the WAIS, the scree gives three factors, a number that does not lead to replicable factors (Gorsuch, 2000).

The suggestion to define the number of factors as the first factor among the trivial roots is what gives three factors for the psychological variables instead of two. This has been controversial in what some would see as extracting one too many factors. That leads to the question of whether extracting too many or too few factors would be more harmful. The

simulation studies have found that extraction of one too many factors seldom does any harm, but extracting one too few distorts the factors that are extracted.

The extraction of an extra factor in the psychological variables leaves the first two with only minor changes and the third factor has two small loadings in the .20s. The box problem is more interesting. The fourth factor brings in a variable not loaded highly by the previous length, height, and width factors: thickness of the edge of the box. The fourth factor loads thickness highly, and also width to some degree. (There is still a factor with width as its major variable.) It seems that boxes in our culture are likely to be stronger if they are wider, a finding that extends the understanding of this example. Even so, the so-called extra factor does not seem to be a handicap in that the first three factors are essentially unchanged.

Simulation studies have generally found the scree test to be one of the better tests. We assume that the scree plots were by someone with training who knew nothing about how many factors were designed into the study, but this information is missing from most articles. (If the scree rater or raters were not blind as to the number of factors, that would invalidate the ratings.) Do note that it is often a choice between several possible screes, and several investigators may come to a different conclusion from the same roots. This suggests that training may be usefully investigated in future simulation studies.

Evaluation Via Plasmodes of Roots-Based Criteria

The three examples being used are plasmodes in the sense that the actual number and nature of the factors are established. The correct number of factors is two, three, and four for the three examples.

Given the correct number of factors and the roots in Table 6.8, it is apparent that both the criteria of roots greater than 1 and the parallel analysis criteria are incorrect two out of three times. The former always treat all roots less than 1 as nonfactors and the latter usually suggests even fewer factors, and yet two of the examples have clear and replicable factors with roots less than 1. And the scree test suggests three factors for the first example, three or four for the second, and three for the third, meaning it is correct for the first two examples but misses the third.

With the different results for the simulation studies compared to the three plasmodes here, what is to be concluded? The most likely conclusion is that the simulations used factors stronger than those found in the last two examples. This suggests that an assumption for the use of parallel analysis is that the factors of interest are assumed to have loadings of .8 or so by at least two or three variables. That may be doable in areas with well-established factors, but that is seldom the case in exploratory factor analyses of little-researched areas.

Two conclusions can be reached. The first is that simulation studies should contain more small factors. The second is that root-based criteria may be a dead end for procedures for establishing the number of factors in EFA. (These conclusions apply to both CA and CFA.)

Residual Based Criteria

The purpose of all models of factor analysis is to reproduce the variables. The better that is done, the better the correlations among the variables and the better the variable scores are reproduced. When the reproduced correlation matrix is subtracted from the observed correlation matrix, the result is referred to as the residual matrix. In the perfect data set with the perfect analysis, all of the residual correlations would be zero. To the degree that the residuals are nonzero, then either another factor is needed or these are the chance variations in the correlations due to sampling error. A number of proposals have been made for basing an index for the number of factors on functions of the residuals.

Although the root tests have been for EFA number of factors, residual-based indices of the adequacy of the factors extracted have also been developed for CCFA. In the case of CCFA, an index is evaluating not only the number of factors (as in EFA), but also the adequacy of the specified factors. Two different hypothesized patterns may produce sufficiently different residuals so that one of the hypothesized patterns is obviously better than the other. Hence, for CCFA the criteria evaluate the total solution.

Statistical Significance

The residual matrix can be tested for significance. If the test is significant, there is more nonrandom variance that can be extracted. If it is nonsignificant, then the extracted factors as a set account for all the correlations among the variables. As with all significance tests, a larger N allows detection of smaller differences.

The psychological variables whose CCFA is presented in Table 6.3 also had a chi-square of 5.53 with df of 8. That has a $p > .10$, so the residual matrix after the hypothesized two factors had been extracted has no covariance that could be considered nonchance. Hence, the conclusion is that these two factors account for all the correlations among these six variables. Note an unusual characteristic of testing the residuals for significant: A nonsignificant result is desirable.

So the problems of predicting a null hypothesis occur, primarily that there are many ways of getting nonsignificant results. These include having variables of low reliability and too small an *N*.

The significance test of the residuals tests whether the extracted factors do account for *everything*. There is no other commonly used test of significance that operates in this manner; all others test whether the hypothesis accounts for some of the variance, not all of it.

The significance test used gives a chi-square. Chi-squares are additive, and two approaches to analyzing the goodness of fit are based on this additivity. First, a suggestion has been to divide the chi-square by the degrees of freedom, giving the average chi-square (which is also *F* because $df = 1$). The advantage of the average chi-square is that it allows a comparison across models that have used a different number of parameters. The averaged chi-square for the six-variable example is .69, because any chi-square/*F* this small shows no chance of anything significant. It further reinforces the conclusion that these two factors are sufficient to account for all the correlations among the six variables.

The second use of chi-square, using the knowledge that chi-squares are additive, notes that the chi-square can be broken down to give a direct comparison between two models when one of the two models is a subset of the other. This is useful because it changes the test from one that tests whether we know everything to one that tests whether adding the hypothesized factor helps. For example, the WAIS began with two factors, Verbal and Performance. And three factors is a solution suggested by the Scree test. Does adding a third and fourth factor account for significantly more of the correlations? That can be tested by running two CCFAs, one for the two factors and one for the four factors (which includes the same parameters for the first two factors as the two-factor model). Each will give a chi-square; the four-factor chi-square is subtracted from the two-factor chi-square to give the chi-square of the two additional factors (the *df* of the difference is computed by subtracting the larger *df* from the smaller). The chi-squares and difference for the WAIS are in Table 6.9. Using the difference chi-square and the difference degrees of freedom allows a significance test of adding the further specification. It does not have the problems of the significance test of residuals, wherein the test is of a null hypothesis. The difference in Table 6.9 is highly significant, showing the four-factor solution to be better significantly than the two-factor solution. But also note that, with the *N* of 1,105, even the four-factor model does not account for all the significant variance. No one has proposed more than four factors because they would be so small that they could not be interpreted. Although a chi-square test has been proposed for EFA, it has seldom been found to be useful.

TABLE 6.9 Tests for the Adequacy of Fit in CCFA: WAIS-III

	Chi-square			RMS	
Model	*df*	Value	Chi/*df*	Residual	Square
Two factors	43	324.4	7.52	.051	.088
Four factors	38	232.1	6.11	.041	.075
Chi-square difference	5	92.3			

Size of the Residuals

Because both EFA and CFA are to reduce the residuals to zero, measuring the size of the residuals is another method of evaluating the adequacy of the factor solution. There are two major approaches, one based on the residuals themselves and another based on the results when they are converted to partial correlations. The former is used with CCFA and the later with EFA.

Two residual-based tests are given in Table 6.9 for the WAIS-III analyses. *RMS* can be interpreted as *root mean square* because it is, roughly, the square root of the mean of the squared residuals. Two varieties of this criterion are in the table (Steiger & Lind, 1980); as can be seen, they generally proceed in the same direction because both are related to the same residuals. By these, it can be seen that the two additional factors do reduce the residuals. (Bentler & Bonett, 1980 give another set of useful indices for CCFA; for overviews of the many indices available for CCFA, see Bentler, 1989.)

An index of the residuals in EFA is Velicer's MAP (minimum averaged partial). Instead of using the residuals, MAP standardizes the residuals by converting them to partial correlations by dividing by the variances of the two variables involved (the residuals are the variances and covariances with the factors partialled out). These are then, in the original MAP, squared and averaged. The logic is that each factor that accounts for covariation among the variables will reduce the residual covariances. As long as the main diagonal elements remain relatively stable, then each factor extracted will lower the averaged partial. But when a factor is extracted that is based less on the covariances, then it will be more specific to one variable and lower the variance (in the main diagonal) of that variable. Because this is divided into the residual covariance, dropping the variance without dropping the covariance increases the partial correlations for that variable. So the minimum averaged partial is used for the number of factors. Minor shifts in MAP suggest that two

solutions are about the same. The principle noted previously that one too many factors is better than one too few suggests that the minimum with the greater number of factors be chosen. MAP is still evolving in that a version that raises the partial to the fourth power (instead of the original second power) is being tried. Evaluative studies suggest it is often helpful (Velicer, Eaton, & Fava, 2000).

The MAPs for the three examples were computed (using the fourth power). For the six psychological variable data, the first three MAPs were .03, .03, and .11, thus giving two factors. For the box data, the five MAPs were .07, .08, .11, .07, and .11, thus suggesting four factors. For the WAIS-III, they were .002, .002, .008, .027, and .061, suggesting two or three factors.

Simulation studies are supportive of MAP in its fourth-power form, but it misses the WAIS factors by suggesting one too few.

How to Select the Number of Factors

The procedures noted previously are typical of the possibilities for establishing the number of factors. Dozens of others have been suggested. As yet, they provide no clear solution to deciding the number of factors. For example, parallel analysis has been one of the best in the simulation studies and yet was clearly inadequate in the plasmode examples used in this chapter. What, then, shall be done?

There are two principles that can guide in establishing the number of factors. First, the prime criterion is the *replication* of the factors. The fact that the WAIS-III four-factor solution has been often replicated in children and adults and in the United States and in Canada is the convincing rationale for the number of factors. What the criteria for the number of factors suggest is much less important than whether the factors can be replicated. The replication of EFA results can occur through a CCFA in a new sample as long as it is accepted that the CCFA will not help in the development of the model, only in its confirmation. More impressive is the confirmation of the EFA factors in new EFA analyses. EFA presents the best possible solution regardless of past results, whereas CCFA analyzes whether the hypothesized solution is one appropriate solution (there could be others, some even better). Both types of confirmation are useful.

The second principle for establishing the number of factors is the interest of the investigator. In the WAIS data, one factor gives *g*, general intelligence, which has been historically of considerable usefulness. Two factors gives the classical Verbal and Performance IQs. And four factors adds two smaller factors that may be of special interest to some investigators, but without rejecting the other two factors.

Consider the three solutions for the box data in Table 6.10. The one-factor solution is technically good. The factor, Volume, accounts for a surprising amount of the variance. It seems that the prime difference among boxes graduate students had available to measure was overall size. The three-factor solution is as expected: length, weight, and height. That also is a good solution. With the four-factor solution, the factors are length, thickness of edge, height, and width. This also could be a useful solution. It depends on the context of the study and the investigator's intent as to which solution is preferable.

In the two examples that can have different numbers of factors extracted, nothing is lost by going to the solution with the greater number of factors. The four-factor box solution still contains length, height, and width factors, and the volume factor occurs at the second-order level. The four-factor WAIS solution still contains verbal and performance types of factors, with *g* occurring at the second-order level.

It appears that taking out more factors and doing a higher-order analysis is the best answer to the number of factors.

TABLE 6.10 Alternate Solutions for the Box Data

	Factor Solutions							
Variable	1 Factor	3 Factor			4 Factor			
1. Length squared	.84	1.02	−.08	−.02	1.07	.01	−.03	−.07
2. Height squared	.85	.05	.17	.81	.05	.15	.83	.03
3. Width squared	.85	.02	.98	.01	−.03	.67	.03	.51
4. Length + width	.96	.73	.46	−.13	.64	.27	−.12	.38
5. Length + height	.96	.69	−.04	.43	.66	−.01	.45	.00
6. Width + height	.96	.18	.58	.41	.11	.77	.12	−.01
Inner diagonals								
7. Longest	.92	.74	.18	.11	.68	.11	.12	.17
8. Shortest	.91	.49	.27	.26	.36	.05	.29	.41
9. Space	.97	.76	.15	.16	.66	.03	.18	.26
10. Edge thickness	.74	.03	.71	.14	.11	.77	.12	−.01

Note. The values greater than 1.0 are because the loadings are weights, not correlations, and the factors have high intercorrelations.

Rotate several different numbers of factors with only causal use of the criteria suggested for the number of factors. Replication will ultimately decide which factors are useful.

My current conclusion is that the appropriate number of factors is, and will be for the immediate future, a semisubjective decision—partially because our attempts to create a universal rule for the number of factors has failed so far. Investigators may well rotate several different numbers of factors and pick the one that they feel is most interpretable, just so long as it has a greater, rather than lesser, number of factors. Indeed, it may be desirable to report the several solutions that replicate. However, this position means that one can never say that one number of factors is the only number that can be, just that it is one of the possible replicable solutions. In the WAIS-III data, one factor gives *g*, two factors give the classical verbal and performance, three factors are not replicable, and four factors give verbal, perceptual organization, working memory, and processing speed. Which solution is best depends on the work at hand, but only the solution with the greater number of factors and a higher-order analysis gives the total story.

Relating Factors

Relating Factors to Other Available Variables

Not all variables that may be available from the sample should be included in a factor analysis. Nonfactored variables may be from another domain or have correlated error with variables being factored (as when scoring the same responses two different ways). How do the factors relate to other data available from the sample but that have not been included in the factor analysis?

There are several major reasons for relating factors to variables not in the factor analysis: Some variables cannot be included in a factor analysis. First, variables that are a linear combination of other variables cannot be included (principal factor and maximum likelihood extraction methods give an infinite number of solutions if a linear combination is included). An example is the total score from a set of items. The total score is a linear combination of the items and so must be excluded. Second, any variable that has correlated error with another variable would adversely affect a factor analysis. One example is scoring the same items for several scales. Another example is including the power of a variable to test for curvilinear relationships, which has correlated error with the original variable. The correlated error can be modeled in a CCFA but not in an exploratory factor analysis. The relationship of factors to total scores, scores that have one or more items in common, and powers of variables can only be analyzed using extension analysis.

Nominal variables cannot be included in a factor analysis, but how the factors relate to such variables may be of interest. Whether the nominal variable be gender, ethnicity, experimental versus control groups, or some other variable, the relationship of nominal variables can be statistically analyzed by extension analysis.

What is the relationship of the factors to ordinal or better variables excluded from the factor analysis? Is a factor related to age or education? Assuming that one is not interested in an age or education factor, it is more appropriate to use extension analysis than to include such variables in the factor analysis.

The need to relate to other variables also occurs when a factor analysis is computed to reduce multiple colinearity or to orthogonalize a set of variables. If the factors are of the independent variables, then those factors need to be entered into the appropriate statistical analysis to relate them to the dependent variables, which were not in the factor analysis. If the dependent variables were factored, then these factors need to be related to the independent variables. If both independent and dependent variables were factored, then the independent variable factors would be tested to see how they correlate with the dependent variable factors.

Another need for extension analysis is in evaluating proposed scales from factor analysis. The factor analysis identifies the dimensions or constructs that can be measured. It also provides the correlations of each item with each factor. Items are then selected for a proposed scale for Factor A from those items that correlate highly with Factor A but not with the other factors. The item set for the scale would contain those that show the highest correlation with the factor—that is, have the highest factor validity. In practice, the first several items for a proposed scale are obvious due to their high correlations. But does adding a moderately correlated item increase or decrease the factor validity of the proposed scale? That question is answered by scoring the items to measure the factor both without and with the moderate item to determine which version of the proposed scale gives the highest factor validity. The set of items with the best factor validity with Factor A is then recommended to be the scale to measure Factor A. (Note that this cut-and-fit item selection method requires a large N to avoid capitalizing on chance, and the observed factor validities will shrink when computed in a new sample. A cross-validation sample is recommended for reporting factor validity correlations.) Relating factors to variables not in the factor analysis is called *extension analysis* because it extends the factors to new variables. The older procedure for extension analysis has been based on computing factor scores (formulas can be used so the actual scores need not be computed), and then analyzing these factor

scores with the extension variables. There are several methods for computing factor scores, but the choice is usually between only two variations. The first is multiple regression analysis. The variables loaded by a factor are the predictors and the factor is the dependent variable. The regression analysis provides the beta weights, which are then used to calculate the factor scores. However, regression weights have the *bouncing beta* problem: Unless the sample is over 400, they bounce around when a new sample is collected or when the variable mix is changed slightly.

The instability of beta weights has led to the other recommended procedure for computing factor scores: unit weighting. Unit weighing is defined as adding together the scores of the variables that have high weights in the multiple regression from the variables to the factors, after the variables have been converted to the same metric, e.g., Z scores. Each of the variables clearly related to the factor is weighted +1 if the weight is positive or −1 if the weight is negative. With samples less than 400, unit weights have higher factor validities when cross-validated than do multiple regression weights (due to the latter's capitalization on chance).

Factor scores have problems. In addition to indeterminacy of CFA scores, each variable weighted in the scoring equation has its unique part added to the score as well as the part loaded by the factor. This is the same problem that occurs when the items are correlated with the total score from the items. The item-total correlations are inflated because that part of the item not measuring the construct is included both in the total score and in the item. To avoid correlations inflated by correlated error, item-remainder correlations have been suggested. Correlating the item with a total score from the remaining items eliminates the inflated correlation. However, it also ignores the valid part of the item that should be part of the total score, and so gives an underestimate of the correlation. The same is true with factor scores: Items or variables contributing to that factor score will have higher correlations due to the shared error.

In the past, extension analysis has been by factor scoring, even when called extension analysis. For that reason it has the problems previously noted for variable–factor score (or item-total and item-remainder) correlations.

However, a new extension analysis procedure has been developed without these problems (Gorsuch, 1997). The new extension analysis can find the effect size and significance levels between factors and any variable collected from the same sample but not in the factor analysis. These may be variables such as gender or age and age squared to check for curvilinear relationships with age. For item development, it gives the factor validity of any proposed scale (without inflation from correlated error).

Extension analysis allows factor analysis to be used as a scoring procedure. The dependent variables (or the independent variables, or both) can be factored and then the other variables of interest related directly to the factors.

Extension analysis is only available at this time in one statistical package (Gorsuch, 1994). However, a detailed example in the original article (Gorsuch, 1997) shows how it can, with patience, be computed even with a hand calculator.

Relating Factors to Prior Studies

Do the factors of Study B replicate those of Study A? This question is addressed by CCFA, which applies when the variables are the same in the two studies. The test is of the overall solution.

But not all situations can be solved by CCFA. What if only part of the factors are included in the new study? Or what if the population sampled is so different that new factors could occur and that would be important information? In these types of situations, some prefer another EFA as a multitailed test that allows unexpected factors to occur. Then it is appropriate to use a factor score procedure. The factor score weights from the first sample are used in the new sample to produce first study factor scores. They are correlated with the new study factors through the Gorsuch extension analysis (not by new study factor scores because they would have correlated error with the first study factor scores and so have inflated correlations). This extension analysis extends the factor analysis of the second study to the factor scores created with the weights from the first study.

The only appropriate measure of how factors relate is how they correlate. (Coefficients of congruence remain a poor choice and cannot be recommended except in rare cases when no estimate of the factor correlations is possible.)

RELEVANT RESEARCH DESIGN PRINCIPLES

The preceding discussion has dealt with the general models and proceedings for factor analysis, whether it be by components or maximum likelihood, exploratory or confirmatory methods. There are, however, some aspects of crucial importance that have not been directly germane to the specifics of the discussion to this point. These are mostly the same issues as in any research study and can be summarized briefly.

The variable and case sampling are crucial to a quality solution. Here is a remainder of aspects to be noted for a factor analysis that hold true of all good research studies:

- Each variable should be interpretable so that a factor's loading or not loading is meaningful.

- The higher the reliability of the variables, the higher the correlations and the communality.
- The higher the validity of the variables, the more meaningful the results.
- For significance testing, uniqueness scores should be normally distributed.

Variables should have similar distributions in the sample for maximum correlations. They need not be normally distributed, but a variable with a skew incompatible with the majority of the other variables should be avoided.

All variables need to have some cases that score high and some that score low. Normal distribution is fine, but it is not desired if it obscures true highs and true lows. This avoids restriction of range, which lowers observed correlations and so weakens the factor structure. The sample size needs to be large enough for stable correlations. Before the plasmode and simulation studies, the best guess was that the N needed would be a function of the number of variables being analyzed. Unlike multiple regression analysis and many previous discussions (e.g., Nunnally, 1967; Gorsuch, 1974, 1983), factor analytic accuracy appears to be relatively independent of the number of variables (with the exception that, for mathematical reasons, the total N must always be larger than the total number of variables). However, both plasmode and simulation studies suggest that the N and the purpose of the study are crucial. The N gives the stability of a correlation, and stability increases as the square root of the N decreases. A zero correlation with an N of 100 has a standard error of .10, 150 is .08, 200 is .07, 300 is .06, and 400 is .05. This is a reasonable guide to sample size. Because the purpose of a study is generally to distinguish between observed correlations of .30 and .40, for example, the safe sample size is 400. If one just wishes to determine which correlations are different from zero and is only interested in correlations .30 and higher, an N of 150 is reasonable. A larger sample is needed for item factor analysis because one needs to differentiate between correlations differing by only .10 ($N = 400$) and to reduce capitalization on chance in item selection.

The number of variables that each factor is expected to load should be in the range of three to six. Fewer than three variables makes a factor difficult to define, so using four to six is better. Simulation studies have suggested more variables be used, but these are only when there are available new variables that are truly different from the original ones, except for being loaded by the same factors. Experience suggests that such a situation seldom occurs, and the variables added after the first six lead to minor factors.

More than six variables can lead to problems due to the sensitivity of factor analysis. In EFA, a factor with more than six variables often gives two subfactors. Unless the factors are restricted to being uncorrelated (in which case there is no recovery), the factor of interest tends to be recovered as a higher-order factor. That higher-order factor may relate well to another analysis which, using fewer variables, finds the factor among the primary factors. In CCFA, more than six variables per factor often leads to statistically significant residuals—even when they are not relevant—due to minor factors found within the six variables.

REFERENCES

Bentler, P. M. (1989). *EQS Structural equations program manual.* Los Angeles: BMDP Statistical Software.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588–606.

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

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: W. B. Saunders Co.

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

Gorsuch, R. L. (1984). Measurement: The boon and bane of investigating religion. *American Psychologist, 39,* 228–236. (Reprinted in *Psychology of religion: Personalities, problems, possibilities,* 1991, pp. 267–283, by H. N. Malony, Ed., Grand Rapids, MI: Baker Book House.)

Gorsuch, R. L. (1990). Common factor analysis versus component analysis: Some well and little known facts. *Multivariate Behavioral Research, 25,* 33–39.

Gorsuch, R. L. (1994). UniMult: For univariate and multivariate data analysis [Computer program and manual]. Pasadena, CA: UniMult.

Gorsuch, R. L. (1997). Exploratory factor analysis: Its role in item analysis. *Journal of Personality Assessment, 68,* 532–560.

Gorsuch, R. L. (1997). New procedure for extension analysis in exploratory factor analysis. *Educational and Psychological Measurement, 57,* 725–740.

Gorsuch, R. L. (2000). Results of the WAIS-III Canadian Study. In D. Wechsler (Ed.), *WAIS-III: Wechsler Intelligence Scale for Adults: Canadian Manual* (3rd ed.). Toronto, Canada: The Psychological Corporation, Harcourt Brace.

Gorsuch, R. L., & Chan, M. Y. A. (1991). *Development and evaluation of a Chinese translation of the State Trait Anxiety Inventory and the Beck Depression Inventory.* Paper presented at the annual meeting of the International Congress of Psychology, San Francisco, CA.

Kaiser, H. F. (1970). A second generation Little Jiffy. *Psychometrika, 35,* 401–415.

Lauhenschlagen, G. J., Lance, C. E., & Flaherty, V. L. (1989). Parallel analysis criteria: Revised equations for estimating the latent roots of random data correlation matrices. *Educational and Psychological Measurement, 49,* 339–345.

Mershon, B., & Gorsuch, R. L. (1988). Number of factors in the personality sphere: Does increase in factors increase predictability of real life criteria? *Journal of Personality and Social Psychology, 55,* 675–680.

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

Snook, S. C., & Gorsuch, R. L. (1989). Component analysis vs. common factor analysis: A Monte Carlo study. *Psychological Bulletin, 106,* 148–154.

Steiger, J. H., & Lind, J. C. (1980, May). *Statistically-based tests for the number of common factors.* Paper presented at the annual Spring Meeting of the Psychometric Society in Iowa City.

Thompson, B. (2000). Q-technique factor analysis: One variation on the two-mode factor analysis of variables. In L. Grimm & P. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 207–226). Washington, DC: American Psychological Association.

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1–28.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes, E. (Eds.), *Problems and solutions in human assessment: A festschrift to Douglas Jackson at seventy,* pp. 47–71. Norwell, MA: Kluwer Academic Publishers.

CHAPTER 7

Clustering and Classification Methods

GLENN W. MILLIGAN AND STEPHEN C. HIRTLE

The purpose of this chapter is to provide a review of the current state of knowledge in the field of clustering and classification as applied in the behavioral sciences. Because of the extensive literature base and the wide range of application areas, no attempt or assurance can be made that all domains of study in this area have been covered. Rather, the main research themes and well-known algorithms are reviewed. In addition, the chapter includes a survey of the issues critical to the analysis of empirical data with recommendations for the applied user.

Clustering and classification methods as discussed here are within a context of exploratory data analysis, as opposed to theory development or confirmation. Some methods or strategies useful for theory confirmation are included as appropriate.

One difficulty in this area is that no unifying theory for clustering is widely accepted. An interesting result in the field of clustering is that the standard statistical assumption of multivariate normality as a basis for the derivation of such algorithms has not automatically led to a superior clustering procedure. Because of derivational difficulties and empirical experience with various approaches, we have today a plethora of methods. Some of these methods work well in certain circumstances, and some of these appear seldom if ever to work as intended. Often, applied users of the methodology are unaware of various issues concerning the performance of clustering and classification methods.

A second problem faced by researchers new to the field is that the literature base is indeed vast and spans virtually all fields of human endeavor. The Classification Society of North America is now in its third decade of publishing an annual bibliographic review called the *Classification Literature Automated Search Service* (Murtagh, 2000). Each issue includes references of upwards of 1,000 scientific articles.

The wide range of application areas creates an additional problem for the applied researcher. Reading scientific articles and textbooks outside of one's own area of expertise can be difficult yet essential to get a good mastery of the topic. Some of the best work in this area has been published in engineering and the biological sciences in addition to outlets normally used by the social sciences community. The reader will see the diversity of disciplines represented in the references section for this chapter. It is useful to note that much of the development of this methodology has appeared in applied journals and less so in the mainstream statistical and mathematical journals.

This chapter continues with a section on data preparation, data models, and representation, including a discussion of distance and similarity measures. Three illustrative applications of classification methods are presented in turn. A section on clustering algorithms covers a wide range of classification methods. In addition, this section includes a discussion of the

recovery performance of clustering methods. The fourth section covers a variety of issues important for applied analyses such as data and variable selection, variable standardization, choosing the number of clusters, and postclassification analysis of the results. The chapter concludes with a section that covers a variety of extensions and issues in classification.

DATA PREPARATION AND REPRESENTATION

The basic data for input to a cluster analysis can consist of either a square or rectangular matrix, with or without replications. For a typical cluster analysis scenario, assume there is a matrix of n objects measured on m features. Depending on the context, the objects have been denoted in the literature as items, subjects, individuals, cases, operational taxonomic units (OTUs), patterns, or profiles, whereas the features have been denoted variables, descriptors, attributes, characters, items, or profiles (Legendre & Legendre, 1998). Thus, the reader of multiple articles must be careful in interpretation, as the same terminology has been used in the literature to refer to both the n rows or the m columns in the data matrix, depending on the specific context of the classification problem.

While is it possible for a cluster-analytic approach to analyze the data in the rows and columns of the rectangular matrix directly, it is more typical first to transform the $n \times m$ rectangular matrix into an $n \times n$ symmetric proximity matrix. Each entry x_{ij} in the transformed matrix represents either similarity of item i to j, in which case we call it a *similarity* matrix, or the dissimilarity of item i to j, in which case we call it *dissimilarity* or *distance* matrix. Alternatively, one could convert the $n \times m$ rectangular matrix to an $m \times m$ symmetric matrix to measure the similarity between features. Sneath and Sokal (1973) denoted the analysis of an $n \times n$ matrix *R analysis,* whereas the analysis of an $m \times m$ matrix was denoted *Q analysis.*

It is also possible to collect similarity or dissimilarity measures directly. For example, Shepard (1963) uses a confusion matrix (Rothkopf, 1957) for the identification of Morse code as an indication of the perceptual similarity of each pair of codes. A matrix entry x_{ab} would indicate how many times the transmitted code for letter a is perceived as letter b. Note that such a matrix would most likely be nonsymmetric. Thus, the researcher would first want to construct a symmetric matrix through the average or weighted average of the two cells x_{ab} and x_{ba}, unless the clustering method explicitly represents asymmetries in the solution (Furnas, 1980; Hirtle, 1987; Okada, 1996).

Carroll and Arabie (1980, 1998) denote the $n \times n$ matrix as *two-way, one-mode* data, whereas the $n \times m$ matrix is referred to as *two-way, two-mode* data. That is, the number of ways reflects the number of dimensions in the data set, while the number of modes reflects the number of conceptual categories represented in the data set. Examples of two-way, one-mode data include confusions, correlations, and similarity ratings (in psychology); frequency of communication between individuals (in sociology); or the subjective distance between locations (in behavioral geography). Examples of two-way, two-mode data include individual responses to questionnaire items (in psychology), n specimens measured on m characteristics (in biology), or ratings of products by consumers (in marketing). This terminology can be extended to include *three-way, two-mode data,* for which two-way, one-mode data is replicated for individual subjects or groups of subjects. Examples of three-way, two-mode data include individual ratings of similarity (in psychology), or the buying patterns of consumer groups (in marketing).

Ultrametric and Additive Inequalities

The results of classification analyses are often represented by tree diagrams, which reflect the inherent relationships in the underlying model. The most common representation is a rooted, valued tree, also called a *dendrogram,* as shown in panel A of Figure 7.1. Here, each node in the tree is joined at a specific height, as indicated by the scale on the right side of the figure. In this case, the set of heights can be shown to satisfy the ultrametric inequality (Johnson, 1967). Specifically, if h_{ij} is the smallest value for which items i and j cluster, then

$$h_{ij} \leq \max(h_{ik}, h_{jk}) \text{ for all } i, j, k.$$

That is, the three heights between each pair of a triple of points can be thought of as an isosceles triangle, with the equal sides being at least as long as the third side.

An alternative tree model is the path-length, or *additive,* tree shown in panel B of Figure 7.1. Here, the dissimilarity

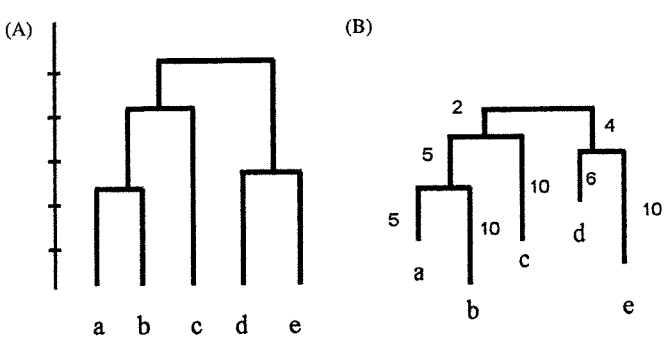

Figure 7.1 Example of two rooted trees. Panel A shows an example of an ultrametric tree, whereas panel B shows an example of a path-length or additive tree.

between items is reflected in length of the paths between the terminal nodes (Buneman, 1971; Corter, 1996; Dobson, 1974). An additive tree is governed by the additive inequality, which states that if d_{xy} is the path length between x and y, then

$$d_{ij} + d_{kl} \leq \max(d_{ik} + d_{jl}, d_{il} + d_{jk}) \text{ for all } i, j, k, l.$$

The ultrametric tree is therefore a special case of the additive tree, where the leaf nodes are all equally distant from the root node. In an additive tree, this restriction does not hold. For example, in the tree shown in panel B of Figure 7.1, $d_{ab} = 15$, $d_{ac} = 20$, and the $d_{bc} = 25$, whereas in panel A the $d_{ac} = d_{bc}$.

In all cases just discussed, only the leaves are explicitly labeled. The researcher may often label the internal nodes on an ad hoc basis to assist in the readability and interpretation of the clusters. The reader, however, should be warned that in such cases the internal labels are arbitrary and not defined by the clustering algorithm.

Classification Data as Tree Models

Corter (1996) argued for the acknowledgment of clustering and trees as models of proximity relationships, rather than as the result of an algorithm for fitting data. The distinction here is subtle but important. Cluster analysis can begin with the notion of some existing underlying clusters. The clusters might be subject to noise and error and vary in dispersion and overlap. The clusters are sampled with measurements taken on a variety of attributes, which are then subjected to a cluster analysis to recover the true clusters. This approach is described in many of the general references in cluster analysis, such as Aldenderfer and Blashfield (1984), Hartigan (1975), or Jain and Dubes (1988).

An alternative framework proposed by Corter (1996) considers the problem of representing a similarity matrix by a structure, such as an additive or ultrametric tree. That is, the information within a matrix has a structure that can alternatively be captured in a representation with fewer parameters than are found in the original data matrix. Pruzansky, Tversky, and Carroll (1982), using this approach, examined the properties of data matrices that would lead to the best fit of spatial or tree representations. Their approach was based on two distinct analyses. First, artificial data were generated by choosing points either randomly from a two-dimensional space or from a randomly generated tree. Noise, at various levels, was then added to some of the data matrices. Not surprisingly, they found that multidimensional scaling algorithms, such as KYST (Kruskal & Wish, 1978), which generated a two-dimensional solution, resulted in a better fit for the spatially generated data, whereas a clustering method, such as ADDTREE (Sattath & Tversky, 1977), resulted in a better fit for the tree-generated data.

The next step was more interesting. Are there patterns in the data matrix that would lead one to adopt one method or the other? As diagnostic measures, they calculated the skewness of the distances and the number of elongated triples. A triple of distances was said to be elongated if the medium distance was closer to the longer distance than to the shorter distance. The analysis by Pruzansky et al. (1982) showed that spatially generated data tended be less skewed and had fewer elongated triples, while the tree-generated data were more negatively skewed and had a larger percentage of elongated triples. As a final step, these diagnostic measures were confirmed using various empirical data sets, which were thought to be best modeled by a tree or by a spatial representation. Thus, for deciding between spatial and tree-based representations, the analyses of Pruzansky et al. (1982) suggest that appropriate diagnostic techniques might suggest which class of models is more appropriate for a given a data set.

EXAMPLES

At this point, it is useful to consider three examples of cluster analysis from the literature. The first example is based on kinship data from Rosenberg and Kim (1975), which has been analyzed in detail by Carroll and Arabie (1983), De Soete and Carroll (1996), and others. The task that the subjects performed in the initial study was to sort kinship terms into any number of piles so that each pile consisted of related terms and there were at least two piles. By taking the total number of times that a subject put two terms in the same pile, one can construct a similarity matrix between terms. Rosenberg and Kim (1975) asked some subjects to sort the terms once, while others were asked to sort the terms multiple times. Using the data matrix from female subjects, De Soete and Carroll (1996) constructed a dendrogram, as shown in Figure 7.2, using a least-squares ultrametric tree-fitting procedure called LSULT (De Soete, 1984). The resulting ultrametric tree representation, which accounts for 96.0% of the variance in the original data matrix, encapsulates the standard anthropological model of kinship terms (Carroll & Arabie, 1983). The tree divides direct kin, such as grandparents, from collaterals, such as cousins. It further divides the direct kin into the immediate family versus ±2 generations. Within these clusters, further groupings occur on the basis of generation (e.g., mother and father are clustered). In this case, there is great benefit in considering the entire representation. That is, if one were to truncate the tree and declare that kin terms are best represented as three clusters or seven clusters, much information would be lost.

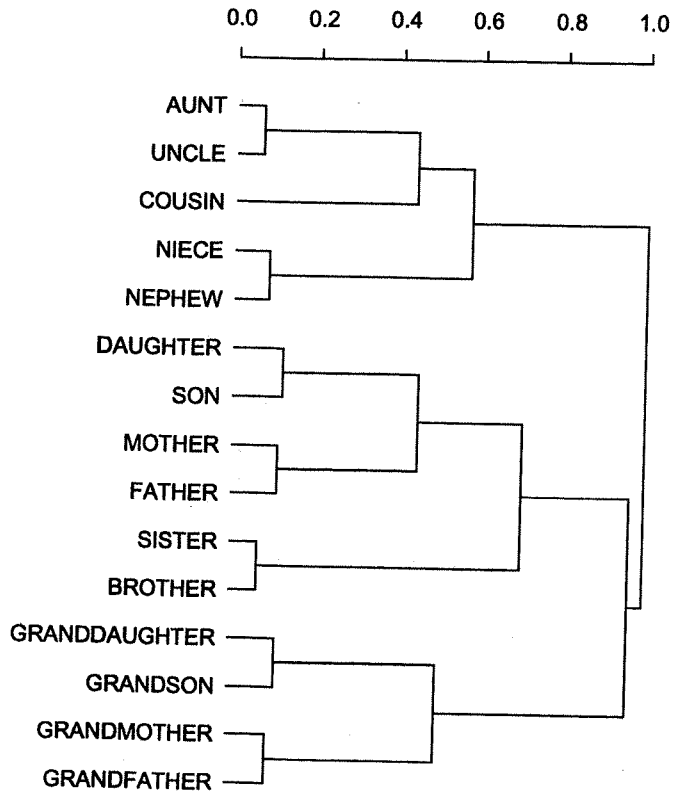

Figure 7.2 Dendrogram for kinship data as produced by a least-squares ultrametric tree-fitting procedure by De Soete and Carroll (1996).

Another example where the entire tree is important is shown in Figure 7.3, which comes from De Soete and Carroll (1996). Figure 7.3 displays an additive tree representation of data collected by Arabie and Rips (1973), based on an earlier study by Henley (1969). In the study, 53 American students were asked to judge the similarity among 30 animals. The representation was generated by LSADT (De Soete, 1984), which is a least-squares additive tree-fitting procedure, and accounts for 87.3% of the variance in the data. As in the previous example, the entire tree representation is interesting, and truncating the tree would be misleading. In addition, some relationships represented by the additive tree would not be represented in an ultrametric tree. For example, *dog* and *cat* are closer to each other in the representation than *tiger* and *wolf*, even though *dog* and *wolf* are in one cluster of canine animals and *cat* and *tiger* are in another cluster of feline animals. An ultrametric representation would force *dog* and *cat* to be the same distance apart as *tiger* and *wolf*, assuming they remained in the canine and feline clusters.

It is also worth emphasizing in both of these examples that only the terminal nodes are labeled. However, implicit labels could be generated for the internal nodes, such as *grandparents* or *felines*. Carroll and Chang (1973) developed one of the few clustering methods for generating a tree representation with labeled internal nodes from a single data set. However, the method has not been widely used, in part because of the limited number of stimulus sets that contain both terminal and nonterminal item names.

One final example is based on a cluster analysis by Lapointe and Legendre (1994). In their study, they produced a classification of 109 single-malt whiskies of Scotland. In particular, the authors of the study were interested in determining the major types of single malts that can be identified on the basis of qualitative characteristics as described in a well known connoisseur's guide (Jackson, 1989). The primary data consisted of 68 binary variables, which represented the presence or absence of a particular descriptive term, such as a *smoky* palate, a *salty* nose, *bronze* in color. The 109×68 matrix was transformed into a 109×109 lower triangular matrix of proximities using the Jaccard (1901) coefficient of similarity, which is based on the number of attributes that a pair of items has in common. The proximity matrix was used to construct the dendrogram using Ward's method, which is described in the next section. The resulting dendrogram in shown in Figure 7.4. In contrast with

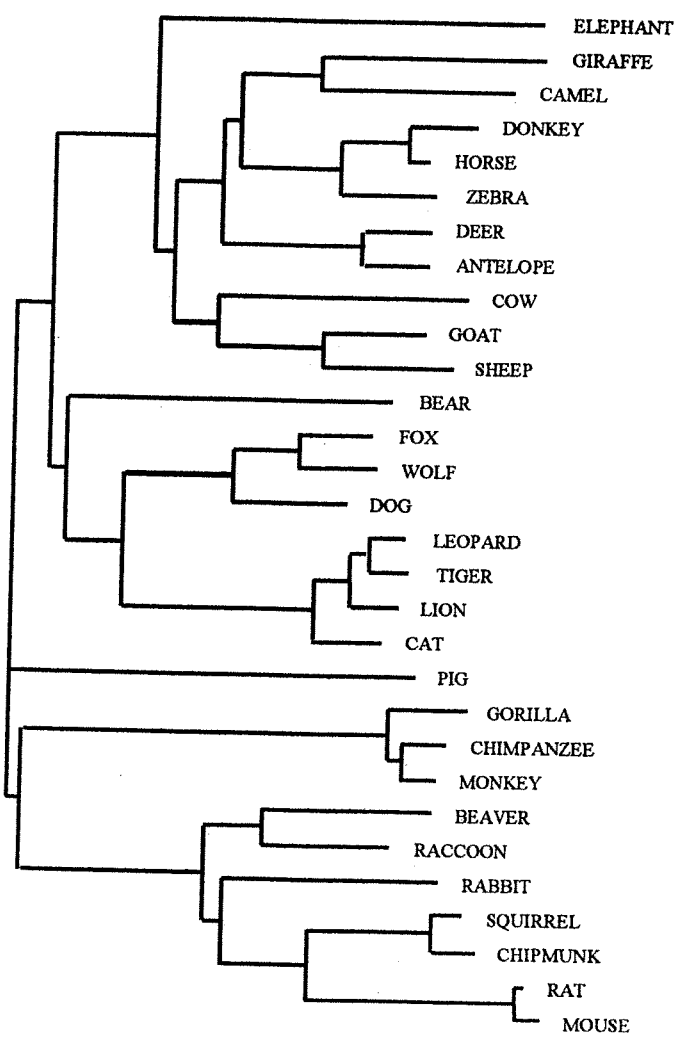

Figure 7.3 Additive tree representation for the animal similarity data as produced by De Soete and Carroll (1996).

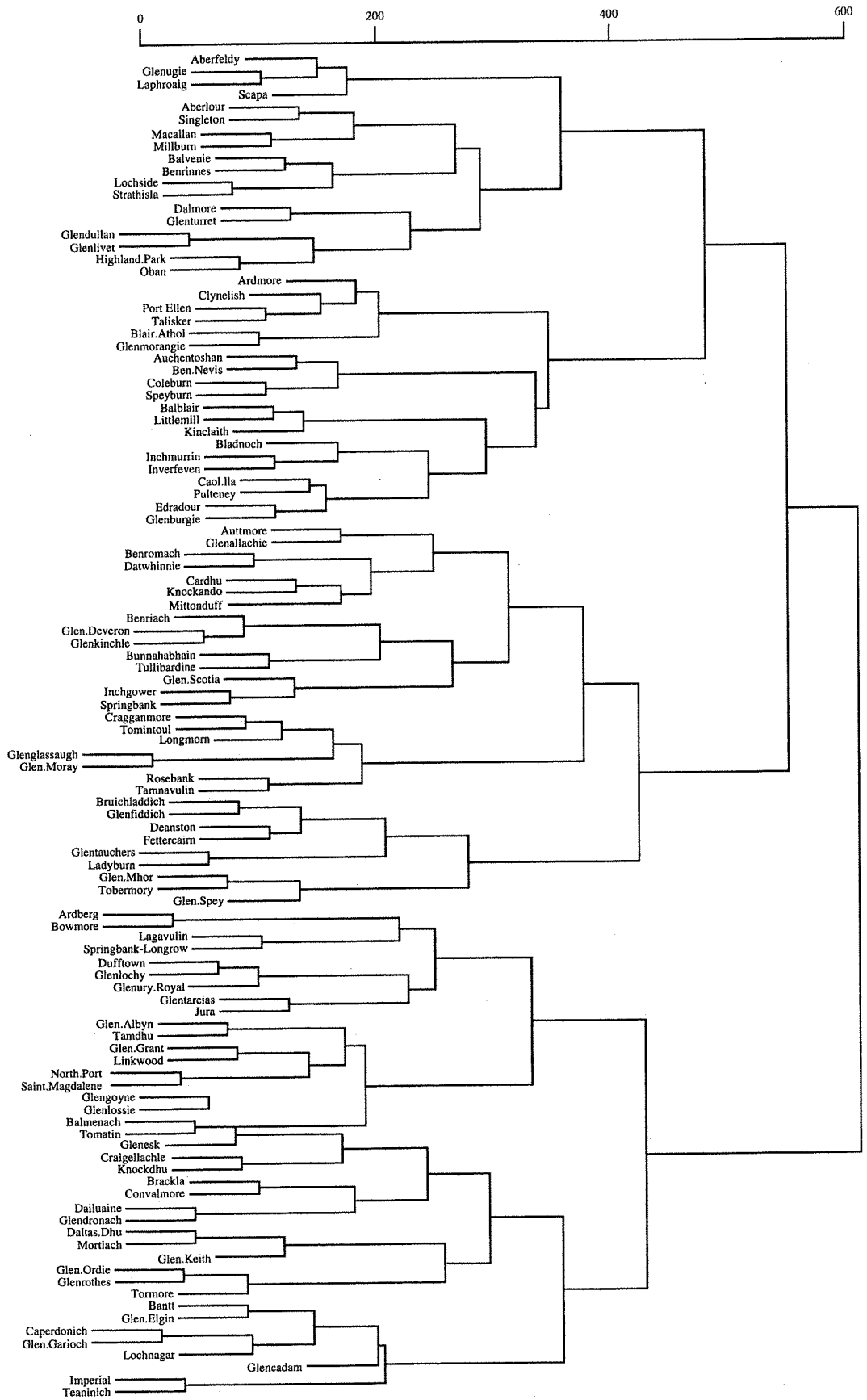

Figure 7.4 The Lapointe and Legendre (1994) classification of single malt scotch whiskies.

the previous examples, the authors are less interested in the structure of the entire tree. Instead, the goal of the study was to identify an unknown number of distinct groups. As a result of the analysis, the dendrogram was truncated to generate 12 identifiable classes of whiskeys, each labeled with a letter of the alphabet in Figure 7.4.

ALGORITHMS

There are several fundamental issues relating to the selection of a suitable clustering algorithm. First, the method must be appropriate for the type of cluster structure that is expected to be present in the data. Different clustering criteria and cluster formation methods yield different types of clusters. Second, the clustering method needs to be effective at recovering the types of cluster structures that it was intended to find. Nearly all clustering methods are heuristics, and there is no guarantee that any heuristic is effective. Finally, software support needs to be available for applied analyses. It is our experience that the latter issue tends to drive method selection with only limited regard for the first two concerns.

For those readers who wish to make a more in-depth study of clustering algorithms, several textbooks and survey articles have been written. These include the texts by Anderberg (1973), Everitt (1993), Gordon (1999), Hartigan (1975), Jain and Dubes (1988), Legendre and Legendre (1998), Lorr (1983), and Späth (1980). Survey articles include Gordon (1987), Milligan and Cooper (1987), and Milligan (1996, 1998). Although some of these sources are more dated than others, they include a wealth of information about the topic.

The next three sections offer a review of the major types of clustering methods that have been proposed in the literature. Included in each section is a discussion concerning the issue of selecting a clustering method appropriate to the type of cluster structure expected to be present in the data. The fourth section reviews the performance of a range of clustering methods in finding the correct clustering in the data.

Agglomerative Algorithms

Agglomerative algorithms are the most common among the standard clustering algorithms found in most statistical packages. Here, each of the n objects is considered to be cluster consisting of a single item. The algorithm then iterates through $n - 1$ steps by combining the most similar pair of existing clusters into a new cluster and associating a height with this newly formed cluster (Gordon, 1996). Different algorithms use different methods for defining the most similar pair, associating a height, and defining a proximity measure

TABLE 7.1 Coefficients to Generate Clustering Techniques Based on the Formalization of Lance & Williams (1966)

Clustering Method	α_i	β	γ
Single link	1/2	0	−1/2
Complete link	1/2	0	1/2
Group-average link	$\frac{n_i}{n_i + n_j}$	0	0
Weighted-average link	1/2	0	0
Centroid	$\frac{n_i}{n_i + n_j}$	$\frac{-n_i n_j}{(n_i + n_j)^2}$	0
Median	1/2	−1/4	0
β-Flexible	$\frac{1-\beta}{2}$	$-1 \le \beta \le 1$	0

between the new cluster and the previously established clusters. In particular, if the new cluster is given by the agglomeration of C_i and C_j, then one can define the new dissimilarities measures by the general formula given by Lance and Williams (1966, 1967) as follows:

$$\begin{aligned} &dissim(C_i \cup C_j, C_k) \\ &\quad = \alpha_i d(C_i, C_k) + \alpha_j d(C_j, C_k) + \beta d(C_i, C_j) \\ &\qquad + \gamma |d(C_i, C_k) - d(C_j, C_k)| \end{aligned}$$

Different choices of the parameters $\{\alpha_i, \alpha_j, \beta, \gamma\}$ define different clustering algorithms as shown in Table 7.1. For example, $\alpha_i = 1/2, \gamma = -1/2$, defines the single-link algorithm where the new dissimilarity coefficient is given by the smallest distance between clusters. This algorithm tends to generate unstable clusters, where small changes in the data matrix result in large changes in the dendrogram (Gordon, 1996). However, it is one of the few clustering algorithms that would be able to detect clusters that are the result of a long chain of points, rather than a densely packed cluster of points.

Complete link clustering corresponds to $\alpha_i = 1/2$, $\gamma = 1/2$. Single and complete link clustering are based solely on the rank order of the entries in the data matrix and thus can be used with ordinal scale data. Most other algorithms require interval scale data. Of the interval scale techniques, group-average link $[\alpha_i = n_i/(n_i + n_j)]$ and weighted-average link $(\alpha_i = 1/2)$ demonstrate greater success at cluster recovery, as shown later in this chapter, than do either of the ordinal scale techniques. Group-average link is also commonly denoted as UPGMA (for unweighted pair group mean average), whereas the weighted average link method is commonly denoted as WPGMA (weighted pair group mean average; Sneath & Sokal, 1973). Additional information on combinatorial clustering methods can be found in Podani (1989).

Divisive Algorithms

For divisive algorithms, the reverse approach from agglomerative algorithms is used. Here, all n objects belong to a single cluster. At each step of the algorithm, one of the existing clusters is divided into two smaller clusters. Given the combinatorial explosion of the number of possible divisions, divisive algorithms must adopt heuristics to reduce the number of alternative splittings that are considered. Such algorithms often stop well before there are only single items in each cluster to minimize the number of computations needed. Still, the problem of finding an optimal division of clusters for several criteria has been shown to be NP-hard (which implies that the computational time will most likely grow exponentially with the size of the problem) for several clustering criteria (Brucker, 1978; Welch, 1982).

Optimization Algorithms

An alternative approach to iterative algorithms is to reconsider the problem by transforming a dissimilarity matrix (d_{ij}) into a matrix (h_{ij}) whose elements satisfy either the ultrametric or the additive inequality. Optimization algorithms have been developed using a least-squares approach (Carroll & Pruzansky, 1980), a branch-and-bound algorithm (Chandon, Lemaire, & Pouget, 1980), and other approximation approaches (Hartigan, 1967). One promising technique was an approach developed by De Soete (1984). The technique, which is discussed later in this chapter, has been successful at addressing the problem of determining optimal weights for the input variables.

Selecting a Clustering Method

This section focuses on the issue of evaluating algorithm performance. One approach commonly used in the literature is the analysis of real-life data sets. It is not unusual for various articles to attempt to establish algorithm performance by using only one or two empirical data sets. Thus, validating a heuristic method is always questionable. In many cases the results are considered valid because they correspond to some general or intuitive perspective. Several criticisms of this approach exist. First, one must recognize that a very small sample size has been used to establish validity. Second, one can always question the author's a priori grouping of the data. Third, how are we to know that clusters actually exist in the empirical data? Few authors consider a null clustering condition. Finally, assuming that clusters are present, how can we determine that the correct cluster structure was found? These criticisms can seldom if ever be addressed properly through the use of empirical data sets for validation purposes.

Most classification researchers have turned to the use of computer-generated data sets for establishing clustering validity. Simulation or Monte Carlo experiments allow the researcher to know the exact cluster structure underlying the data. This strategy has the advantage that the true clustering is known. The extent to which any given clustering algorithm has recovered this structure can be determined. Because of the use of artificially generated data sets, simulation results can be based on hundreds or thousands of data sets. Thus, sample size is not an issue.

There is a serious weakness in the use of simulation methods. In every case, such results are limited on the basis of generalizability. That is, the Monte Carlo results may be valid only for the types of cluster structures and distributions that were present in the generated data sets. Thus, the effectiveness of the algorithms may not extend to other data structures that are possible in applied analyses. Thus, it is important to establish replicability of simulation results from differing studies. It is especially valuable when different researchers achieve similar results using different strategies for data generation and evaluation. Such replications offer investigators more confidence in the selection of methods for applied analyses.

In terms of results on the recovery of underlying cluster structure, agglomerative hierarchical algorithms have been the most extensively studied. Three reviews of Monte Carlo clustering studies covering various time frames were published by Milligan (1981a), Milligan and Cooper (1987), and Milligan (1996). The validation studies have examined a number of factors that might affect recovery of the underlying clusters. Many studies have included an error-free data condition. The clustering present in the error-free data typically was so distinct that almost any method should have been able to perform well with this sort of simple and obvious data structure. Clustering methods that fail with error-free data would not be suitable for most applied research settings.

A second factor examined has been the introduction of some sort of error, either on the underlying variables or directly to the similarity measures. This condition has the capability of being tuned to a gradient of increasing noise. An effective clustering method should be capable of finding clusters that have been hidden by moderate amounts of error in the data.

A different sort of error involves the introduction of outlying data points to a core set of elements that defines a suitable cluster structure. Unusual observations are not unusual in behavioral research. A clustering method used for applied analyses should have some insensitivity to the presence of such data points.

The population distribution used to conceptualize and generate the clusters themselves need not be multivariate normal. Nonnormality may be present in many empirical data sets, and a clustering method should be able to recover well-defined clusters in such circumstances. Furthermore, alternative population distributions serve to generalize the Monte Carlo results. Few simulation studies have included more than one type of distribution. The generalization exists across different studies using differing underlying population probability models.

The number of clusters in the underlying data can be varied easily and, thus, can serve to ensure that a given clustering method is not sensitive to this factor. The clustering method adopted should not have differential effectiveness on this factor.

The relative sample size of clusters can be systematically varied as well. Some clustering methods do not respond properly to the presence of unequal cluster sizes. This is not a desirable result, and it has implications for applied analyses. The characteristic can be demonstrated most easily by generating data sets with varying cluster sizes.

Some authors have varied the number of variables that are used to construct the artificial data. Since the data are first transformed to a similarity measure, most clustering methods do not directly analyze the original data. However, the number of variables may influence the information captured by the similarity measure and, hence, influence the method's ability to recover the underlying clusters. Other factors have been included in one or more studies. These include the use of more than one similarity measure for the data and the number of underlying dimensions from a principal component representation of the variable space, among others.

Simulation results for a set of hierarchical methods are presented first. Validation results for five such methods are reported in Table 7.2, adapted from Milligan and Cooper (1987). It is important not to overinterpret the results in the

TABLE 7.2 Monte Carlo Validation Results for Hierarchical Methods

	Method				
Study	Single Link	Complete Link	Group Average	Ward's Method	Beta Flexible
Baker (1974)					
Low error	.605	.968			
Medium error	.298	.766			
High error	.079	.347			
Kuiper & Fisher (1975)					
Medium size	.579	.742	.710	.767	
Five clusters	.444	.690	.630	.707	
Unequal sizes	.663	.705	.702	.689	
Blashfield (1976)	.06	.42	.17	.77	
Mojena (1977)	.369	.637	.596	.840	
Mezzich (1978)					
Correlation	.625	.973			
Euclidean	.648	.943			
Edelbrock (1979)					
Correlation	.90	.80	.96		
Euclidean	.62	.63	.70	.88	
Milligan & Isaac (1980)	.30	.64	.70	.57	
Bayne, Beauchanp, Begovich, & Kane (1980)					
Configuration 1	.53	.68	.66	.70	
Configuration 2	.55	.76	.75	.76	
Edelbrock & McLaughlin (1980)					
Correlation	.858	.813	.880		
Euclidean	.690	.780	.858	.873	
Milligan (1980)					
Zero error	.974	.995	.998	.987	.997
Low error	.902	.970	.997	.989	.994
High error	.777	.880	.948	.940	.945
Scheibler & Schneider (1985)					
Correlation	.43	.49	.81	.78	.73
Euclidean	.04	.38	.16	.79	.77

Note. For details on the nature of the recovery values, see Milligan and Cooper (1987).

table because the recovery index is not the same across all studies. Direct numerical comparisons should be made within a given study, and not across different experiments. The measures do have the common characteristic that recovery performance improves as the index approaches 1.00, which indicates perfect cluster recovery.

The simulation results in Table 7.2 contain some important lessons for the applied user. In most cases, there appears to be an advantage in favor of Ward's (1963) method and the β-flexible approach. Performing somewhat more erratically, the group-average method can be competitive as gauged by cluster recovery, but not always. The effectiveness of the β-flexible approach from these studies led to some improvements on this method by Milligan (1989a) and Belbin, Faith, and Milligan (1992).

A particularly important result seen in Table 7.2 is that the single-link method has consistently performed poorly, even in the case of error-free data where distinct clustering exists. Furthermore, single link is especially sensitive to most any form of error added to the data. Cheng and Milligan (1995a, 1996a) also demonstrated that the single-link method was remarkably sensitive to outliers present in the data. That is, the method can be adversely affected by the presence of only one outlier. An outlier in a clustering context refers to an entity that does not fall within the general region of any cluster. Although some authors have argued that the method possesses optimal theoretical properties (e.g., Fisher & Van Ness 1971; Jardine & Sibson; 1971), simulation and empirical evidence suggest that this is an unsuitable method for most applied research.

Simulation-based research on nonhierarchical partitioning methods has not been as extensive as for the hierarchical routines. *K*-means (MacQueen, 1967) algorithms have been the most frequently examined methods to date. Simulation results for such methods are presented in Table 7.3. Generally, these studies were based on error-free data sets. The simulation-based literature indicates that the recovery performance of some partitioning methods can be competitive with those found for the best hierarchical procedures. As before, the reader is warned not to overinterpret the numerical recovery values between studies as they are based on different indices.

Most of the generated data sets used to establish the results in Table 7.3 were multivariate normal and should have been the ideal application context for the normal theory-based clustering methods such as the Friedman and Rubin (1967) and Wolfe's (1970) NORMIX procedures. Unfortunately, such methods performed inconsistently in these studies. Less sophisticated methods, such as *k*-means algorithms, can produce equivalent or superior recovery of cluster structure.

TABLE 7.3 Monte Carlo Validation Results for Nonhierarchical Clustering Methods

Clustering Method	Average Recovery	Recovery With Rational Seeds
Blashfield (1977)		
Forgy *k*-means	.585	
Convergent *k*-means	.638	
CLUSTAN *k*-means	.706	.643
Friedman-Rubin trace **W**	.545	
Friedman-Rubin |**W**|	.705	
MIKCA trace **W**	.560	
MIKCA |**W**|	.699	
Mezzich (1978)		
Convergent *k*-means: correlation	.955	
Convergent *k*-means: Euclidean distances	.989	
Ball-Hall ISODATA	.977	
Friedman-Rubin |**W**|	.966	
Wolfe NORMIX	.443	
Bayne et al. (1980)		
Convergent *k*-means	.83	
Friedman-Rubin trace **W**	.82	
Friedman-Rubin |**W**|	.82	
Wolfe NORMIX	.70	
Milligan (1980): Low error condition		
MacQueen's *k*-means	.884	.934
Forgy's *k*-means	.909	.996
Jancey's *k*-means	.926	.993
Convergent *k*-means	.901	.996
Scheibler & Schneider (1985)		
CLUSTAN *k*-means	.67	.78
Späth's *k*-means	.55	.77

Note. Average recovery for *k*-means methods corresponds to random starting seeds. "Rational Seeds" were centroids obtained from Ward's or group-average methods.

One characteristic discovered from the set of studies reported in Table 7.2 concerns the nature of the cluster seeds used to start the *k*-means algorithms. The *k*-means algorithms appear to have differential recovery performance depending on the quality of the initial configuration. This effect was systematically studied by Milligan (1980). The results reported by Milligan indicated that starting seeds based on randomly selected sample points were less effective than was the use of rational starting configurations. Rational starting seeds markedly improved the recovery performance of all *k*-means methods. In light of these results, Milligan and Sokol (1980) proposed a two-stage clustering algorithm that was designed to improve the recovery of the underlying clusters. Subsequently, other researchers have endorsed this approach or developed useful refinements (see Punj & Stewart, 1983; Wong, 1982; Wong & Lane, 1983).

Overall, more research on the comparative evaluation of clustering methods is needed. We have good information on certain types of methods. However, for other methods or approaches the current knowledge base on algorithm

performance is weak or badly lacking. For example, there have been a number of recent developments. An interesting approach to clustering, called MCLUST, has been proposed by Raftery, Fraley, and associates (see Fraley & Raftery, 1998). To date, an independent evaluation of this approach has not been published.

STEPS IN A CLUSTER ANALYSIS

A fundamental principle in classification is that as the level of error increases in the data, or in the specification of one or more factors relating to the clustering, the ability to recover the underlying cluster structure is reduced. Thus, a number of issues must be addressed while conducting an applied analysis in addition to the choice of clustering method.

Sometimes these decisions are not apparent to the researcher. For example, a researcher may select a clustering software package that makes one or more of these decisions without user intervention. The researcher should be alert to the fact that these decisions were made and that they directly affect the quality of the clustering results.

When applied research is published using clustering methodology, we recommend that the specific actions taken during the classification process be clearly articulated. This practice is essential to allow subsequent researchers the ability to evaluate, compare, and extend the results. Examples abound in the literature where authors have failed to provide such information (see Milligan, 1996). Critical information includes the choice of similarity measure, the clustering algorithm used to form the groups, the determination of the number of clusters, and information on the sample and variables used in the analysis.

Several key elements or decision points in the clustering process are reviewed in this section. Best practical suggestions, based on the current state of knowledge, are offered. These suggestions relate to the selection of the elements to be clustered, the selection of the variables to cluster, issues concerning variable standardization, the selection of the number of clusters, and the validation of empirical analyses.

Selecting the Data Set

The issue of selecting the data elements in a cluster analysis has seen limited research. This issue is critical because it is the sample of data elements selected for study that define the resulting cluster structure. Several fairly simple principles can guide the researcher. Unlike traditional inference-based statistical procedures, random samples are not required for an effective cluster analysis. Certainly, the selected sample should accurately represent the underlying clusters, but not necessarily in proportion to their size in the larger population. In the absence of this consideration, it is likely that small population segments may not be detected in a cluster analysis. Oversampling these small populations would likely serve to enhance their recovery in the cluster analysis. Furthermore, some clustering methods have some bias to find clusters of relatively equal size, and this tendency can be used to good advantage.

Of course, random sampling would be desirable if it is essential for the researcher to be able to generalize the results of the study to a target population. However, doing so would imply a more theoretically driven analysis as opposed to a more exploratory study. Random or stratified sampling would be useful in replication studies or in more advanced studies attempting to validate a contextual theory.

The selection of the sample elements should consider the overall size of the database. A second sample or a split-half sample would be helpful for validation purposes, as discussed later in this chapter. As suggested by Milligan (1996), one possible approach is to place artificially generated ideal-type individuals or subjects in the data set. The researcher specifies the values for each variable of an ideal-type individual. The ideal type would represent a subject or other experimental object that would represent the norm for each group or cluster suspected to be present in the data. One or possibly more ideal types would be specified for each hypothesized cluster. The presence of the correct ideal type or types in a cluster would support the researcher's conceptualization for the hypothesized clustering. On the other hand, if markedly different ideal types appear in the same cluster, then the researcher's theory or the cluster analysis is suspect. The presence of clusters without ideal types may represent groups not yet defined by the researcher's theory, or possibly subgroups of a larger cluster. The user should be warned that the use of ideal types is a temporary process. The presence of ideal types in the final clustering may change the assignment of other elements in the data set. The relative influence of individual data elements has been explored by Cheng and Milligan (1995a, 1995b, 1996a, 1996b).

Related to the issue of influential data points is the issue of outliers. Outliers in a clustering context deserve special consideration. As stated previously, an outlier in a clustering context refers to an entity that does not fall within the general region of any cluster. Note that outliers may or may not have influence on the clustering solution obtained, and some data points near or in a cluster may have an influential effect on the clustering process.

An early simulation study on the effect of outliers in clustering was conducted by Milligan (1980). This research

confirmed that as the percentage of outliers increased, the ability of hierarchical clustering methods to recover the underlying structure decreased. Some methods were less affected than others. More recent results concerning the effect of outliers on hierarchical methods can be found in Milligan (1989a) and Belbin et al. (1992). This more recent research suggests that Ward's (1963) method may not be as seriously affected by the presence of outliers as first suspected. Similarly, Belbin et al. (1992) demonstrated desirable characteristics with respect to outliers for two versions of the β-flexible method. Overall, the impact of outliers appears to be less severe for *k*-means methods.

The applied user of clustering methodology can adopt several different strategies for dealing with outliers. One can eliminate those elements that appear to be outliers to the overall set of data. Alternatively, the relationship between the obtained clusters and the suspected outliers can be investigated after an initial clustering is completed. A third alternative is to use a clustering method resistant to the presence of outliers. Selected parameterizations of the β-flexible hierarchical clustering procedure and Ward's (1963) minimum variance method may be good selections, as well as some of the *k*-means algorithms.

Variable Selection and Weighting

Clustering methods differ profoundly from traditional statistical inference models. Standard statistical requirements such as the assumption of normally distributed data generally do not apply within the clustering framework. That is, the methods are heuristics, and they were often developed without consideration of an underlying probability model for the data.

Another common misconception is that the presence of correlated variables in the data set is somehow bad or undesirable. Researchers often fail to realize that the correlations among variables may be a result of the natural cluster structure in the data. Attempts to eliminate these correlations would likely serve to distort or hide the structure in the data. Numerous applied analyses have attempted to eliminate intervariable correlation by means of principal components or other multivariate methods. Unfortunately, the routine application of principal components or other factoring techniques prior to clustering is appropriate only in those cases where the clusters are hypothesized to exist in the factor space and not in the original data. Sneath (1980) has shown that clusters embedded in a high-dimensional variable space may not be correctly identified in a reduced number of orthogonal components.

A different issue relates to the selection of variables to include in the cluster analysis. Care must be exercised in selection of the variables. Most reference works in the clustering area fail to offer strong advice on this issue. Only those variables that are believed to help discriminate among the clusters in the data should be included in the analysis. Far too many analyses have been conducted by including every available variable. Some users have gone to great efforts to collect just one more variable without considering its ability to help find the underlying clustering. Instead, the bias should be *not* to include the variable without additional information.

The difficulty in using all available data can result from the added irrelevant variables' serving to mask whatever actual clustering is present in a reduced number of variables. In fact, the addition of only one or two irrelevant variables can dramatically interfere with cluster recovery. Milligan (1980) was the first to demonstrate this effect. In this study only one or two random noise variables were added to data sets where a strong and distinct clustering was present in a reduced set of variables. Fowlkes and Mallows (1983) introduced the term *masking variables,* which is a good description of the effect. Results from the Milligan (1980) study are presented in Table 7.4.

As can be seen in Table 7.4, cluster recovery quickly degraded with even one random noise dimension added to the core data containing distinct clustering. A second dimension continued to diminish the ability to find the true structure in the data. The core dimensions defined a strong clustering in the data. Clearly, there are important implications for applied analyses. The inclusion of just one irrelevant variable may serve to mask or hide the real clustering in the data. It would

TABLE 7.4 Results From Milligan (1980): Mean Recovery Values With Masking Variables

Clustering Method	Error-Free Data	1-Dimensional Noise	2-Dimensional Noise
Hierarchical			
Single link	.974	.899	.843
Complete link	.995	.859	.827
Group average (UPGMA)	.998	.930	.903
Weighted average (WPGMA)	.994	.917	.885
Centroid (UPGMC)	.983	.808	.616
Median (WPGMC)	.976	.808	.661
Ward's method	.987	.881	.855
β-flexible	.997	.904	.863
Average link in cluster	.985	.870	.834
Minimum total *SS*	.935	.837	.780
Minimum average *SS*	.993	.900	.865
Partitioning			
MacQueen's *k*-means	.884	.793	.769
Forgy's *k*-means	.932	.844	.794
Jancey's *k*-means	.927	.867	.823
Convergent *k*-means	.903	.849	.787

Note. Average within-cell standard deviation is .108 and was based on 108 data sets.

be wise to provide a justification for each variable included in the clustering process. The bias should be toward exclusion in the case where doubt exists as to whether the variable may contain information regarding the clustering in the data.

Fortunately, a significant contribution on the problem of masking variables has been made. If Euclidean distances are used with a hierarchical clustering method, then the optimal variable weighting method of De Soete (1986, 1988) may offer helpful protection against masking variables. De Soete's method computes optimal weights for the distance equation:

$$d_{ij} = \left[\sum_{k=1}^{nv} w_k (x_{ik} - x_{jk})^2 \right]^{.5}.$$

The derivation and computation of the weights are complex, and the reader is referred to the work of De Soete (1986, 1988) and Makarenkov and Legendre (2000) for further details. Originally, De Soete's procedure was not intended to detect masking variables. Rather, the purpose was to optimize the fit of the computed distances to an ultrametric structure. The application to masking variables was suggested by one of the example analyses conducted by De Soete (1986). Milligan (1989b) pursued this application and found evidence that the method was effective at dealing with the masking problem. Makarenkov and Legendre (2000) recently have replicated the results concerning the effectiveness of the weights against masking variables. In addition, their work provides an important extension to k-means methods.

The results in Table 7.5 are from Milligan's (1989b) study of De Soete's algorithm. The study compared the recovery performance using equal variable weights to that obtained using optimal weights. As can be seen in the table, recovery performance was greatly enhanced, even when three masking variables were added to the core cluster dimensions. Further research revealed that De Soete's algorithm was assigning effectively zero weights to the masking variables, thus eliminating their noise contribution to the distance computation.

There have been other attempts to deal with the problem of optimal variable weighting. For example, DeSarbo, Carroll, and Green (1984) proposed a procedure called SYNCLUS. The algorithm uses a nonhierarchical k-means method in the clustering process. To date, there has not been a systematic validation study conducted on the SYNCLUS algorithm. Green, Carmone, and Kim (1990) reported that the starting configuration used for the k-means method appears to be a critical factor for the success of the effectiveness of the variable weighting method. Other approaches to the masking problem do not attempt to provide differential weighting of variables. Rather, the method of Fowlkes, Gnanadesikan, and Kettenring (1988) attempts to include or exclude variables in a manner analogous to that used in stepwise regression.

Variable Standardization

With respect to variable standardization, we again find that applied researchers bring potentially ill-advised biases to the clustering process. First, many researchers assume that variable standardization is required in order to prepare the data for clustering. They assert that variable standardization is necessary when the variances among variables differ to any significant degree. Similarly, some authors will argue that standardization is essential when substantial differences exist in the numerical magnitude of the mean of the variables. Otherwise, it is believed that those variables with the larger scales or variances will have an undue influence on the cluster analysis.

Many researchers fail to consider that if the cluster structure actually exists in the original variable space, then standardization can distort or hide the clustering present in the data. Again, as with principal components, standardization would be appropriate if the clusters were believed to exist in

TABLE 7.5 Results From Milligan (1989b): Mean Recovery for Masking Variables Using De Soete's (1988) Variable Weighting Algorithm

	1 Dimension		2 Dimensions		3 Dimensions	
Clustering Method	Equal Weights	Weighted	Equal Weights	Weighted	Equal Weights	Weighted
β-flexible = −.5	.750	.966	.673	.952	.601	.948
β-flexible = −.25	.788	.979	.716	.962	.657	.961
Single link	.812	.883	.647	.840	.473	.820
Complete link	.668	.977	.595	.955	.555	.930
Group average	.859	.980	.809	.965	.732	.957
Ward's method	.764	.968	.675	.955	.627	.947
Column standard deviation	.263	.128	.295	.163	.307	.180

Note. Each mean was based on 108 data sets.

the transformed variable space. This result was first demonstrated in a simple example by Fleiss and Zubin (1969). Other discussions on this topic appeared in Sneath and Sokal (1973) and in Anderberg (1973).

A different bias brought to the analysis by applied researchers is an assumption as to the form of variable standardization to be used. Researchers with a social science or statistics background often assume that variable standardization would be based on the traditional z score:

$$z_1 = \frac{x - \bar{x}}{s}.$$

It turns out that there are number of other ways in which to standardize data so that the influence of variance and relative numerical values can be controlled. Milligan and Cooper (1988) documented several other approaches to variable standardization:

$$z_2 = \frac{x}{s},$$

$$z_3 = \frac{x}{Max(x)},$$

$$z_4 = \frac{x}{Max(x) - Min(x)},$$

$$z_5 = \frac{x - Min(x)}{Max(x) - Min(x)},$$

$$z_6 = \frac{x}{\sum x},$$

and $$z_7 = Rank(x).$$

Milligan and Cooper (1988) evaluated the performance of the various forms of standardization in a large-scale simulation study. Included were the traditional z score (z_1), z_2 through z_7, as well as the unstandardized data represented by z_0 in their study.

Selected simulation results from the Milligan and Cooper (1988) article are presented in Tables 7.6 and 7.7. Each entry in the tables represents the average obtained from 864 data sets. Note that the rows in the tables correspond to the various forms of standardization. The columns in Table 7.6 represent different types of artificially generated data structures. The entries are averages across four clustering methods. Table 7.7 presents similar information broken down by clustering method.

The asterisk notation is unique to these tables and requires explanation. An asterisk indicates that the corresponding standardization method was in the statistically equivalent superior group for a given column. This was, in effect, a test of simple main effects in a factorial ANOVA design. Thus, the asterisk indicates the best performing methods for each condition. Across the conditions explored in Milligan and Cooper (1988), the only standardization procedures that were in the superior group in every case were those methods that standardized by range, namely z_4 and z_5. The consistency of the results was unexpected. Since the publication of the 1988 study, anecdotal evidence reported by numerous researchers has supported the Milligan and Cooper results. Recently, Mirkin (2000) has been developing a mathematical theory as to why standardization by range has been consistently effective. Mirkin and other researchers are likely to continue with this line of inquiry.

TABLE 7.6 Results From Milligan & Cooper (1988): Effect of Standardization Procedure and Alternative Data Structures

Standardization Formula	Separation Level		Maximum Variance Ratio		Global Variance
	Near	Distant	16	100	Experiment
z_0	.662	.821	.745	.739	.621(L)
z_1 & z_2	.672	.837	.755	.754	.936
z_3	.689*	.854*	.771*	.772*	.984*
z_4 & z_5	.693*	.864*	.778*	.780*	.968*
z_6	.674*	.836	.757	.753	.981*
z_7	.639(L)	.768(L)	.693(L)	.713(L)	.839
Overall	.674	.835	.754	.756	.888

Note. The asterisk indicates membership in the statistically equivalent superior group. (L) indicates that the procedure performed significantly worse than the other methods.

Selecting the Number of Clusters

The next significant problem faced in the analysis is the determination of the number of clusters to be used in the final solution. Some clustering methods, such as k-means, require the user to specify the number of groups ahead of time. Other methods require the researcher to sort through and select

TABLE 7.7 Results From Milligan & Cooper (1988): Effect of Standardization Procedure and Clustering Method

Standardization Formula	Clustering Method			
	Single Link	Complete Link	Group Average	Ward's Method
z_0	.608*	.750	.811	.798(L)
z_1 & z_2	.577	.778	.800	.864*
z_3	.622*	.793*	.835*	.836
z_4 & z_5	.609*	.815*	.839*	.851*
z_6	.616*	.761	.813	.828
z_7	.494(L)	.730(L)	.810	.781(L)
Overall	.589	.777	.819	.834

Note. The asterisk indicates membership in the statistically equivalent superior group. (L) indicates that the procedure performed significantly worse than the other methods.

from a sequence of different clustering solutions. This is the case when hierarchical algorithms are selected and the purpose is to find a coherent grouping of the data elements as opposed to a tree representation.

Numerous methods have been proposed for selecting the number of clusters, especially in a hierarchical context. As with many aspects of the clustering process, theoretical developments on this problem have been limited to date. Rather, we have a set of ad hoc methods. The formulas are sometimes called stopping rules for hierarchical clustering methods. The most comprehensive study on the selection of a suitable stopping rule in a hierarchical context is the article by Milligan and Cooper (1985). These authors conducted a comparative evaluation of 30 stopping rules within a simulation framework. The authors considered only those rules that were independent of the clustering method. The generated data sets used by Milligan and Cooper (1985) consisted of error-free structure with distinct clustering. Despite the pronounced clustering present in the data, the results of their study revealed that there was a wide range in the effectiveness of the stopping rules. Selected results from the Milligan and Cooper (1985) study are presented in Table 7.8. The reader is referred to the 1985 article for more detailed performance information and for references for each stopping rule.

The results in Table 7.8 indicate the number of times that a given stopping rule selected the correct number of clusters in the data. The maximum performance rate that could be obtained for any specific number of clusters was 108, and 432 overall. The results in the table include a number of well known approaches such as Mojena's (1977) method, Beale's (1969) pseudo F test, and the rule developed by Calinski and Harabasz (1974). As one reaches the least effective methods at the bottom of the table, the chance selection rate for each cluster level is around 9.

Certainly, more research in the area of stopping rules is needed. The Milligan and Cooper results are from one simulation study, and the potential limitation of generalizability is an important consideration. Independent validation of the performance of the rules with other types of simulated data needs to be undertaken. The reader is warned not to take the performance ranking of the stopping rules as an absolute finding. The rankings produced by Milligan and Cooper (1985) may have been a result of the specific characteristics of the simulated data sets. On the other hand, one might argue that those stopping rules found in the upper third of those tested by Milligan and Cooper might be replicated to some degree in an independent study. Similarly, it would seem unlikely that the least effective rules in their report would perform with a degree of distinction in a different experiment. Support for this conjecture was found by Cooper and Milligan (1988) in a related experiment. In this experiment, the data were subjected to various levels of error perturbation. Although the performance of the rules declined as expected, the relative ranking of the stopping was sustained in the experiment.

TABLE 7.8 Results From Milligan & Cooper (1985): Stopping Rule Performance

	Number of True Clusters				
Stopping Rule	2	3	4	5	Overall
1. Calinski & Harabasz	96	95	97	102	390
2. Duda & Hart	77	101	103	107	388
3. C-index	71	89	91	96	347
4. Gamma	74	86	83	96	339
5. Beale	57	87	95	92	331
6. Cubic clustering criterion	67	88	82	84	321
7. Point-biserial	94	83	66	65	308
8. $G(+)$	52	70	79	96	297
9. Mojena	20	84	93	92	289
10. Davies & Bouldin	54	72	72	89	287
11. Stepsize	96	56	53	68	273
12. Likelihood ratio	64	72	64	68	268
13. $\lvert\log(p)\rvert$	78	71	45	43	237
14. Sneath	34	51	66	83	234
15. Frey & Van Groenewoud	0	76	79	77	232
16. $\log(SSB/SSW)$	0	104	42	66	212
17. Tau	85	77	30	10	202
18. $\bar{c}/\sqrt{k}$	88	80	25	7	200
19. $n\log(\lvert\mathbf{W}\rvert/\lvert\mathbf{T}\rvert)$	0	104	32	13	149
20. $k^2\,\lvert\mathbf{W}\rvert$	0	104	15	27	146
21. Bock	74	15	31	22	142
22. Ball & Hall	0	104	23	1	128
23. Trace Cov(**W**)	0	104	17	0	121
24. Trace **W**	0	104	16	0	120
25. Lingoes & Cooper	37	30	17	16	100
26. Trace $\mathbf{W}^{-1}\mathbf{B}$	0	52	23	9	84
27. Generalized distance	5	22	11	9	47
28. McClain & Rao	9	5	5	6	25
29. Mountford	1	6	1	2	10
30. $\lvert\mathbf{W}\rvert/\lvert\mathbf{T}\rvert$	0	0	0	0	0

For applied analyses, it is recommended that one use two or three of the better performing rules from the Milligan and Cooper (1985) study. The *Statistical Analysis System* (SAS) (Gilmore, 1999) has implemented several of these rules as clustering software options. When consistent results are obtained from the rules, evidence exists for the selection of the specified number of clusters. If partial agreement is found, the user might opt for the larger number of clusters. In this case, one may have an incomplete clustering of the data where two or more groups still need to be merged. Their characteristics will appear to be fairly similar when the researcher is attempting to interpret each cluster. Finally, if no consistency can be found among the rules, the researcher is facing one of several possibilities. Of course, the stopping rules might have failed on the empirical data set at hand. A different outcome is that there is no cluster structure inherent in the

data set. Since most clustering routines will produce a partition (or set of partitions) for any data set, a researcher might assume that there is a significant clustering present in the data. This belief induces a bias against a null hypothesis of no significant clustering in the data in empirical research.

Validation of the Clustering Results

Once the clustering results are obtained, the process of validating the resulting grouping begins. Several strategies or techniques can assist in the validation process. This section covers the topics of interpretation, graphical methods, hypothesis testing, and replication analysis.

Interpretation

An empirical classification will contribute to the knowledge of a scientific domain only if it can be interpreted substantively. To begin the evaluation process, descriptive statistics should be computed for each cluster. The descriptive values can be computed both on those variables used to form the clusters as well as on exogenous variables not involved in computing the clusters. The descriptive information can reveal important differences and similarities between clusters, and it can indicate the degree of cohesiveness within clusters. Skinner (1978) refers to such characteristics as level (cluster mean or centroid), scatter (variability), and shape (covariances and distribution of data within clusters). Similarly, if ideal type markers were used in the analysis, their cluster assignments can be examined for interpretive information.

A different approach is to use a block diagonal matrix display (Anderberg, 1973; Duffy & Quiroz, 1991). Although this technique results in a matrix of numbers, the display approaches that of a graphical presentation. The process is based on rearranging the similarity matrix according to the groups obtained by the cluster analysis. The rows and columns are reordered to place elements in the same cluster in consecutive order. The result is ideally a block diagonal matrix where within-block values represent within-cluster distances or similarities. Entries outside of the blocks correspond to between-cluster distances. If distinct clusters have been recovered by the clustering method, the within-block values should be distinctly different in magnitude when compared to those between blocks. The permuted matrix can be converted to a graphical display if the cells or blocks are shaded according to some rule based on the values of the similarity measures.

A variety of graphical displays have been proposed in the classification literature. For example, Andrews (1972) proposed a bivariate plot where data from a high-dimensional variable space are transformed by means of selected transcendental functions. Andrews argued that similar elements should produce similar transformed profiles in the plot. Bailey and Dubes (1982) developed a different type of display called a cluster validity profile. The profiles were intended to allow for the evaluation of the relative isolation and compactness of each individual cluster. Kleiner and Hartigan (1981) presented a set of graphical methods based on natural-appearing "trees" and "castles." These displays are best suited to hierarchical clustering results. An excellent discussion on the use of graphical methods in a clustering context is found in Jain and Dubes (1988).

Hypothesis Testing

Hypothesis testing is possible in a cluster-analytic situation, but it can be tricky and full of pitfalls for the unsuspecting user. Most testing procedures have been developed to determine whether a significant cluster structure has been found. Because clustering algorithms yield partitions, applied researchers who see such results tend to assume that there must be clusters in their data. However, clustering methods will yield partitions even for random noise data lacking structure.

There are some significant limitations in the use of traditional hypothesis-testing methods. Perhaps the most tempting strategy, given the context of the analysis, is to use an ANOVA, MANOVA, or discriminant analysis directly on the variables that were used to determine the clustering. The partitions obtained from the cluster analysis are used to define the groups for the ANOVA or discriminant analysis. An attempt is made to determine whether there are significant differences between the clusters. Unfortunately, such an analysis is invalid. Since the groups were defined by partitions on each variable, an ANOVA or discriminant analysis will almost always return significant results regardless of the structure in the data, even for random noise. The fundamental problem is that one does not have random assignment to the groups independent of the values on the variables in the analysis. This result was noted by Dubes and Jain (1979) and by Milligan and Mahajan (1980). It is unfortunate that many textbooks on clustering do not emphasize this limitation.

There is a way to conduct a valid inference process in a clustering context. Valid testing procedures take on one of several different approaches. The first approach is called an external analysis, and the test is based on variables not used in the cluster analysis. The second approach is called an internal analysis and is based on information used in the clustering process. These two approaches are considered in turn.

External criterion analysis can be performed using standard parametric procedures. One can test directly for significant

differences between clusters on variables that were not used in the cluster analysis. It is critical for the validity of the test that the variable not be used in forming the clusters.

A different type of external analysis is based on a data partition generated independently of the data set at hand. The partition can be specified from a theoretical model or obtained from a clustering of a separate data set. Hubert and Baker (1977) developed a method to test for the significance of similarity between the two sets of partitions. The test is based on an assumption of independent assignments to groups in the two partition sets. It is important to note that the Hubert and Baker method cannot be applied to two clusterings of the *same* data set. Doing so would not result in two independent groupings of the objects in the study.

An internal criterion analysis is based on information obtained from within the clustering process. These analyses are based on measures that attempt to represent in some form the goodness of fit between the input data and the resulting cluster partitions. There are numerous ways in which to measure the goodness of fit. Milligan (1981b) conducted a study of 30 internal criterion indices for cluster analysis. For an extended discussion of such indices, see Milligan (1981b). Milligan's research indicated that indices such as the gamma, *C*-index, and tau measures should make an effective measure of internal consistency.

The advantage to identifying an effective internal criterion index is that it can serve as a test statistic in a hypothesis-testing context. The test can be used to determine whether a significant clustering exists in the data. The main problem with this approach is the specification of a suitable sampling distribution for the test statistic under the null hypothesis of no cluster structure. One can use randomization methods, or bootstrapping, to generate an approximate sampling distribution. Milligan and Sokol (1980), Begovich and Kane (1982), and Good (1982) have all proposed tests based on this strategy. Unfortunately, software support for this form of testing is not widely available.

Replication Analysis

Replication analysis within a clustering context appears to have been developed by McIntyre and Blashfield (1980) and by Morey, Blashfield, and Skinner (1983). Replication analysis is analogous to a cross-validation procedure in multiple regression. The logic behind replication analysis is that if an underlying clustering exists in the data set, then one should be able to replicate these results in a second sample from the same source and set of variables. There are six steps in a replication analysis. First, one obtains two samples. This can be done by taking a random split-half reliability of a larger data set. Data must be obtained on the same set of variables in both samples. Second, the first sample is subjected to the planned cluster process. Once the clusters have been identified, the cluster centroids are computed from the first sample. These centroids are used in the next step. Third, the distances between the data points in the second sample to the centroids obtained from the first sample are computed. Fourth, each element in the second sample is assigned to the nearest centroid determined from the first sample. This produces a clustering of the second sample based on the cluster characteristics of the first sample. Fifth, the second sample is subjected to the same cluster process as used for the first sample. Note that we now have two clusterings of the second sample. One was obtained from the nearest centroid assignment process, the second from a direct clustering of the data. The final step is to compute a measure of partition agreement between the two clusterings of the second sample. The kappa statistic or the Hubert and Arabie (1985) corrected Rand index can serve as the measure of agreement. The resulting statistic indicates the level of agreement between the two partitions and reflects on the stability of the clustering in the data from two samples.

Breckenridge (1989, 1993) extended this approach to replication analysis and provided performance information on the effectiveness of the approach. The results reported in Table 7.9 are from his 1993 simulation study. The column labeled "Recovery" indicates the degree of agreement between the true cluster assignments and the partitions obtained from the clustering procedure. The column for "Replication" indicates the degree of agreement between the direct clustering of the second sample and the nearest centroid grouping for the same sample. Recall that this grouping was based on the classification from the first sample. The results indicate that the replication means were close in value to average recovery for each method. This finding suggests that replication analysis can be used as a validation tool for applied cluster analysis.

Breckenridge (1993) also reported that replication analysis can be used to help determine the number of clusters in the

TABLE 7.9 Results From Breckenridge (1993): Mean Recovery and Replication Values for Error-Free Data

Clustering Method	Recovery	Replication
β-flexible = −.5	.773	.750
β-flexible = −.25	.761	.738
Single link	.440	.350
Complete link	.695	.654
Group average	.751	.740
Ward's method	.787	.766
Hartigan & Wong *k*-means: (Ward's method seed points)	.785	.797

Note. Averages based on 960 data sets.

TABLE 7.10 Results From Breckenridge (1993): Number of Clusters Selected by the Scree Test

True Number of Clusters	Number Chosen							
	2	3	4	5	6	7	8	9
2	58	1	0	0	0	0	0	0
3	10	48	1	0	0	0	0	0
4	1	12	45	2	1	0	0	0
5	1	2	11	35	9	1	1	0
6	1	2	5	11	32	5	3	1
7	1	0	6	8	12	22	7	4
8	0	1	0	6	13	8	12	20
9	1	5	5	7	5	5	3	29

TABLE 7.11 Number of Source Articles in Psychology Journals on Clustering and Classification by Subdiscipline for 1999

Subdiscipline	*N*	%
Social/personality	28	21.5%
Cognitive/experimental	22	16.9%
Applied/organizational	16	12.3%
General	16	12.3%
Methodological	12	9.2%
Counseling	9	6.9%
Developmental	8	6.2%
Clinical	7	5.4%
Educational	6	4.6%
Neuroscience	5	3.8%

data. Results for 480 data sets are presented in Table 7.10. The columns of the table indicate the number of clusters selected by a scree test, and the rows represent the correct number of clusters in the data. Thus, the diagonal of the table corresponds to the correct specification of the number of clusters identified by the scree test. A scree test is a graphical method used for visually identifying the change in level of a statistical measure. In this application the replication values are plotted across the number of groups in the clustering solution. A notable change in level of the statistic may indicate that the correct number of clusters has been found. The replication scree test was able to specify the correct number of clusters in 58% of the cases. When including those cases that were accurate to within ± 1 cluster, 82% of the data sets were resolved correctly. Thus, further development of the replication methodology seems warranted.

DISCUSSION AND EXTENSIONS

The recommendations presented in this chapter are simply guidelines and not hard and fast rules in clustering. The authors would not be surprised if an empirical data set can be found for each case that would provide a counterexample to the suggested guidelines. Since the classification area is quite active and new research continues to appear, applied researchers are encouraged to review more recent results as time progresses. The journals listed as references for this chapter can serve as a basis for following the current literature. There is no doubt that further advances will reshape our knowledge with respect to this methodology.

Use of Clustering in Psychology and Related Fields

Clustering continues to be used heavily in psychology and related fields. The 1994–1999 editions of the SERVICE bibliographic database list 830 entries in the psychological journals alone. Primary areas of application include personality inventories (e.g., Lorr & Strack, 1994), educational styles (e.g., Swanson, 1995), organizational structures (e.g., Viswesvaran, Schmidt, & Deshpande, 1994), and semantic networks (e.g., Storms, Van Mechelen, & De Boeck, 1994). Table 7.11 lists the 130 articles in psychology journals by subdiscipline for the publication year of 1999, as listed in the SERVICE bibliography. One can note that the subdiscipline list in Table 7.11 spans most of psychology with a remarkably even distribution. In addition, although a number of articles about clustering appear in methodological journals, this category represents only 9% of the publications about clustering and classification. Thus, clustering and classification research remains very healthy in psychology with both methodological developments and substantive applications appearing within the literature on a regular basis.

In addition to research within the mainstream psychology journals, there is a large body of psychological research using classification techniques in several closely related areas. Some of the notable areas include environmental geography, where cluster analysis is used to identify neighborhood structures (Hirtle, 1995); information retrieval, where clustering is used to identify groups of related documents (Rasmussen, 1992); marketing, where there remains a close relationship between data analysis techniques and theoretical developments (Arabie & Daws, 1988); social network theory (Wasserman & Faust, 1994); and evolutionary trees (Sokal, 1985). Arabie and Hubert (1996) emphasize the last three areas as particularly notable for their active use of clustering and for their methodological advances. Psychologists with an interest in the development or novel adaptation of clustering technique are urged to look toward these fields for significant advances.

Relationship to Data Mining

With a recent explosion of interest in data mining, there has also been a resurgence of interest in clustering and classification. Data mining applies a variety of automated and statistical

tools to the problem of extracting knowledge from large databases. The classification methods used in data mining are more typically applied to problems of supervised learning. In such cases, a training set of preclassified exemplars is used to build a classification model. For example, one might have data on high- and low-risk credit applicants. Such problems are well suited for decision trees or neural network models (Salzberg, 1997). In contrast, unsupervised classification is closer to the topic of this chapter in that a large number of cases are divided into a small set of groups, segments, or partitions, based on the similarity across some n-dimensional attribute space. Data-mining problems can be extremely large, with as many as a half million cases in the case of astronomical data (e.g., Fayyad, Piatetsky-Shapiro, Smyth, & Uthurusamy, 1996) or pharmacological data (e.g., Weinstein et al., 1997). Thus, the use of efficient algorithms based on heuristic approaches may replace more accurate, but inefficient, algorithms discussed previously in this chapter.

Han and Kamber (2000) reviewed extensions and variants of basic clustering methods for data mining, including partitioning, hierarchical, and model-based clustering methods. Recent extensions of k-means partitioning algorithms for large data sets include three related methods, PAM (Kaufman & Rousseeuw, 1987), CLARA (Kaufman & Rousseeuw, 1990), and CLARANS (Ng & Han, 1994), which are based on building clusters around medoids, which are representative objects for the clusters. Extensions to hierarchical methods for large databases include BIRCH (Zhang, Ramakrishnan, & Linvy, 1996) and CHAMELEON (Karypis, Han, & Kumar, 1999), both of which use a multiphase approach to finding clusters. For example, in CHAMELEON, objects are divided into a relatively large number of small subclusters, which are then combined using an agglomerative algorithm. Other data-mining clustering techniques, such as CLIQUE (Agrawal, Gehrke, Gunopulos, & Raghavan, 1998), are based on projections into lower dimensional spaces that can improve the ability to detect clusters. CLIQUE partitions the space into nonoverlapping rectangular units and then examines those units for dense collections of objects. Han and Kambar (2000) argued that the strengths of this method are that it scales linearly with the size of the input data and at the same time is insensitive to the order of the input. However, the accuracy of the method may suffer as a result of the simplicity of the algorithm, which is an inherent problem of data-mining techniques.

Software Considerations

Applied researchers may face significant problems of access to user-friendly software for classification, especially for recent advances and cutting-edge techniques. Commercially available statistical packages can seldom keep up with advances in a developing discipline. This observation is especially true when the methodology is not part of the mainstream statistical tradition. It is unfortunate that research-oriented faculty are not able to provide a greater degree of applied software support. Fortunately, the Internet can facilitate access to the research software that is available. For example, the Classification Society of North America maintains a Web site that provides access to an extensive set of software programs that have been made freely available to the research community. The site can be located at http://www.pitt.edu/~csna/. The Web site also provides useful links to commercial software packages, some of which are not widely known. More generally, a wealth of information on the classification community can be found at the Web site.

We still believe that the best advice is for graduate students to develop some skill in writing code in at least one higher level language to support their research activities. In some situations you may just have to write it yourself in order to get the analysis done. One option, among several, is to gain skill at writing macros for the S-Plus (1999) software package. This software package provides a fairly flexible system for handling, manipulating, and processing statistical data.

REFERENCES

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM SIGMOD Conference on Management of Data* (pp. 94–105). New York: Association for Computing Machinery.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis.* Thousand Oaks, CA: Sage.

Anderberg, M. R. (1973). *Cluster analysis for applications.* New York: Academic Press.

Andrews, D. F. (1972). Plots of high-dimensional data. *Biometrics, 28,* 125–136.

Arabie, P., & Daws, J. (1988). The interface among data analysis, marketing, and knowledge representation. In W. Gaul & M. Schader (Eds.), *Data, expert knowledge and decisions* (pp. 10–15). Heidelberg, Germany: Springer-Verlag.

Arabie, P., & Hubert, L. (1996). Advances in cluster analysis relevant to marketing research. In W. Gaul & D. Pfeifer (Eds.), *From data to knowledge, studies in classification, data analysis, and knowledge organization* (pp. 3–16). New York: Springer.

Arabie, P., & Rips, L. (1973). *A 3-way data set of similarities between Henley's 30 animals.* Unpublished manuscript, Stanford University.

Bailey, T. A., & Dubes, R. (1982). Cluster validity profiles. *Pattern Recognition, 15,* 61–83.

Baker, F. B. (1974). Stability of two hierarchical grouping techniques. Case I: Sensitivity to data errors. *Journal of the American Statistical Association, 69,* 440–445.

Bayne, C. K., Beauchamp, J. J., Begovich, C. L., & Kane, V. E. (1980). Monte Carlo comparisons of selected clustering procedures. *Pattern Recognition, 12,* 51–62.

Beale, E. M. L. (1969). *Cluster analysis.* London: Scientific Control Systems.

Begovich, C. L., & Kane, V. E. (1982). Estimating the number of groups and group membership using simulation cluster analysis. *Pattern Recognition, 15,* 335–342.

Belbin, L., Faith, D., & Milligan, G. (1992). A comparison of two approaches to beta-flexible clustering. *Multivariate Behavioral Research, 27,* 417–433.

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four hierarchical agglomerative methods. *Psychological Bulletin, 3,* 377–388.

Blashfield, R. K. (1977). *A consumer report on cluster analysis software. Vol. 3: Iterative partitioning methods* (NSF grant DCR No. 74-20007). State College: Pennsylvania State University, Department of Psychology.

Breckenridge, J. N. (1989). Replicating cluster analysis: Method, consistency, and validity. *Multivariate Behavioral Research, 24,* 147–161.

Breckenridge, J. N. (1993, June). *Validating cluster analysis: Consistent replication and symmetry.* Paper presented at the meeting of the Classification Society of North America, University of Pittsburgh, PA.

Brucker, P. (1978). On the complexity of clustering problems. In R. Henn, B. Korte, & W. Oettli (Eds.), *Optimization and operations research* (pp. 45–54). Berlin: Springer-Verlag.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. In D. Kendall & P. Tautu (Eds.), *Mathematics in archeological and historical science* (pp. 387–395). Edinburgh, Scotland: Edinburgh University Press.

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics, 3,* 1–27.

Carroll, J. D., & Arabie, P. (1980). Multidimensional scaling. *Annual Review of Psychology, 31,* 607–649.

Carroll, J. D., & Arabie, P. (1983). INDCLUS: An individual differences generalization of the ADCLUS model and the MAPCLUS algorithm. *Psychometrika, 48,* 157–169.

Carroll, J. D., & Arabie, P. (1998). Multidimensional scaling. In M. H. Birnbaum (Ed.), *Handbook of perception and cognition* (Vol. 3, pp. 179–250). San Diego, CA: Academic Press.

Carroll, J. D., & Chang, J. J. (1973). A method for fitting a class of hierarchical tree structure models to dissimilarities data and its application to some "body parts" data of Miller's. *Proceedings of the 81st Annual Convention of the American Psychological Association,* 8, 1097–1098.

Carroll, J. D., & Pruzansky, S. (1980). Discrete and hybrid scaling methods. In E. D. Lantermann & H. Feger (Eds.), *Similarity and choice* (pp. 108–139). Bern, Switzerland: Huber.

Chandon, J. L., Lemaire, J., & Pouget, J. (1980). Construction de l'ultrametrique la plus proche d'une dissimilarité au sens des moindres carrés (Fitting a least-squares ultrametric that is nearest to a given dissimilarity). *Recherche Opérationnelle, 14,* 157–170.

Cheng, R., & Milligan, G. W. (1995a). Mapping influence regions in hierarchical clustering. *Multivariate Behavioral Research, 30,* 547–576.

Cheng, R., & Milligan, G. W. (1995b). Hierarchical clustering algorithms with influence detection. *Educational and Psychological Measurement, 55,* 237–244.

Cheng, R., & Milligan, G. W. (1996a). Measuring the influence of individual data points in a cluster analysis. *Journal of Classification, 13,* 315–335.

Cheng, R., & Milligan, G. W. (1996b). *K*-Means clustering methods with influence detection. *Educational and Psychological Measurement, 56,* 833–838.

Cooper, M. C., & Milligan, G. W. (1988). The effect of measurement error on determining the number of clusters in cluster analysis. In W. Gaul & M. Schader (Eds.), *Data, expert knowledge and decisions* (pp. 319–328). Berlin: Springer-Verlag.

Corter, J. E. (1996). *Tree models of similarity and association.* Thousand Oaks, CA: Sage.

DeSarbo, W. S., Carroll, J. D., & Green, P. E. (1984). Synthesized clustering: A method for amalgamating alternative clustering bases with different weighting of variables. *Psychometrika, 49,* 57–78.

De Soete, G. (1984). A least squares algorithm for fitting an ultrametric tree to a dissimilarity matrix. *Pattern Recognition Letters, 2,* 133–137.

De Soete, G. (1986). Optimal variable weighting for ultrametric and additive tree clustering. *Quality and Quantity, 20,* 169–180.

De Soete, G. (1988). OVWTRE: A program for optimal variable weighting for ultrametric and additive tree fitting. *Journal of Classification, 5,* 101–104.

De Soete, G., & Carroll, J. D. (1996). Tree and other network models for representing proximity data. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 157–198). London: World Scientific Press.

Dobson A. J. (1974). Unrooted trees for numerical taxonomy. *Journal of Applied Probability, 11,* 32–42.

Dubes, R., & Jain, A. K. (1979). Validity studies in clustering methodologies. *Pattern Recognition, 11,* 235–254.

Duffy, D. E., & Quiroz, A. J. (1991). A permutation-based algorithm for block clustering. *Journal of Classification, 6,* 65–91.

Edelbrock, C. (1979). Comparing the accuracy of hierarchical clustering algorithms: The problem of classifying everybody. *Multivariate Behavioral Research, 14,* 367–384.

Edelbrock, C., & McLaughlin, B. (1980). Hierarchical cluster analysis using intraclass correlations: A mixture model study. *Multivariate Behavioral Research, 15,* 299–318.

Everitt. B. S. (1993). *Cluster analysis* (3rd ed.). New York: Wiley.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). *Advances in knowledge discovery and data mining.* Cambridge, MA: AAAI/MIT Press.

Fisher, L., & Van Ness, J. W. (1971). Admissible clustering procedures. *Biometrika, 58,* 91–104.

Fleiss, J. L., & Zubin, J. (1969). On the methods and theory of clustering. *Multivariate Behavioral Research, 4,* 235–250.

Fowlkes, E. B., Gnanadesikan, R., & Kettenring, J. R. (1988). Variable selection in clustering. *Journal of Classification, 5,* 205–228.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association, 78,* 553–584.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method?: Answers via Model-Based Cluster Analysis. *Computer Journal, 41,* 578–588.

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association, 62,* 1159–1178.

Furnas, G. W. (1980). *Objects and their features: The metric representation of two class data.* Unpublished doctoral dissertation, Stanford University, Palo Alto, CA.

Gilmore, J. (1999). *Painless Windows: A handbook for SAS users.* Cary, NC: SAS Institute.

Good, I. J. (1982). An index of separateness of clusters and a permutation test for its statistical significance. *Journal of Statistical Computing and Simulation, 15,* 81–84.

Gordon, A. D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A, 150,* 119–137.

Gordon, A. D. (1996). Hierarchical classification. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 65–121). River Edge, NJ: World Scientific Press.

Gordon, A. D. (1999). *Classification: Methods for the exploratory analysis of multivariate data* (2nd ed.). London: Chapman & Hall.

Green, P. E., Carmone, F. J., & Kim, J. (1990). A preliminary study of optimal variable weighting in *k*-means clustering. *Journal of Classification, 7,* 271–285.

Han, J., & Kamber, M. (2000). *Data mining: Concepts and techniques.* San Francisco, CA: Morgan Kaufman.

Hartigan, J. A. (1967). Representation of similarity matrices by trees. *Journal of the American Statistical Association, 62,* 1140–1158.

Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

Hartigan, J., & Wong, M. (1979). A K-means clustering algorithm. *Applied Statistics, 28,* 100–108.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Mathematical Psychology, 8,* 176–184.

Hirtle, S. C. (1987). On the classification of recall strategies using lattice-theoretic measures. *Journal of Classification, 4,* 227–242.

Hirtle, S. C. (1995). Representational structures for cognitive space: Trees, ordered trees, and semi-lattices. In A. V. Frank & W. Kuhn (Eds.), *Spatial information theory: A theoretical basis for GIS.* Berlin: Springer-Verlag.

Hubert, L. J., & Arabie, P. (1985). Comparing partitions. *Journal of Classification, 2,* 193–218.

Hubert, L. J., & Baker, F. B. (1977). The comparison and fitting of given classification schemes. *Journal of Mathematical Psychology, 16,* 233–253.

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

Jardine, N., & Sibson, R. (1971). *Mathematical taxonomy.* New York: Wiley.

Karypis, G., Han, E., & Kumar, V. (1999). CHAMELEON: A hierarchical clustering algorithm using dynamic modeling. *Computer, 32,* 68–75.

Kaufman, L., & Rousseeuw, P. J. (1987). Clustering by means of medoids. In Y. Dodge (Ed.), *Statistical data analysis based on the L 1 Norm* (pp. 405–416). New York: North Holland/Elsevier Science.

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis. New York: Wiley.

Kleiner, B., & Hartigan, J. A. (1981). Representing points in many dimensions by trees and castles (with comments and rejoinder). *Journal of the American Statistical Association, 76,* 260–276.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling.* Thousand Oaks, CA: Sage.

Kuiper, F. K., & Fisher, L. (1975). A Monte Carlo comparison of six clustering procedures. *Biometrics, 31,* 777–783.

Lance, G. N., & Williams, W. T. (1966). A generalized sorting strategy for computer classifications. *Nature, 212,* 218.

Lance, G. N., & Williams, W. T. (1967). A general theory for classificatory sorting strategies I. Hierarchical systems. *Computer Journal, 9,* 373–380.

Legendre, P., & Legendre, L. (1998). *Numerical ecology* (2nd ed.). Amsterdam: Elsevier Science.

Lorr, M. (1983). *Cluster analysis for social sciences.* San Francisco: Jossey-Bass.

Lorr, M., & Strack, S. (1994). Personality profiles of police candidates. *Journal of Clinical Psychology, 50,* 200–207.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.

Makarenkov, V., & Legendre, P. (2000). *Optimal variable weighting for ultrametric and additive tree clustering and K-means partitioning: Method and software.* Manuscript submitted for publication.

McIntyre, R. M., & Blashfield, R. K. (1980). A nearest-centroid technique for evaluating the minimum-variance clustering procedure. *Multivariate Behavioral Research, 15,* 225–238.

Mezzich. J. (1978). Evaluating clustering methods for psychiatric-diagnosis. *Biological Psychiatry, 13,* 265–346.

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

Milligan, G. W. (1981a). A review of Monte Carlo tests of cluster analysis. *Multivariate Behavioral Research, 16,* 379–407.

Milligan, G. W. (1981b). A Monte Carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika, 46,* 187–199.

Milligan, G. W. (1989a). A study of the beta-flexible clustering method. *Multivariate Behavioral Research, 24,* 163–176.

Milligan, G. W. (1989b). A validation study of a variable weighting algorithm for cluster analysis. *Journal of Classification, 6,* 53–71.

Milligan, G. W. (1996). Clustering validation: Results and implications for applied analyses. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 345–379). River Edge, NJ: World Scientific Press.

Milligan, G. W. (1998). Cluster analysis. In S. Kotz, C. Read, & D. Banks (Eds.), *Encyclopedia of statistical sciences* (Vol. 2, pp. 120–125). New York: Wiley.

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.

Milligan, G. W., & Cooper, M. C. (1987). Methodological review: Clustering methods. *Applied Psychological Measurement, 11,* 329–354.

Milligan, G. W., & Cooper, M. C. (1988). A study of variable standardization. *Journal of Classification, 5,* 181–204.

Milligan, G. W., & Isaac, P. (1980). The validation of four ultrametric clustering algorithms. *Pattern Recognition, 12,* 41–50.

Milligan, G. W., & Mahajan, V. (1980). A note on procedures for testing the quality of a clustering of a set of objects. *Decision Sciences, 11,* 669–677.

Milligan, G. W., & Sokol, L. M. (1980). A two-stage clustering algorithm with robust recovery characteristics. *Educational and Psychological Measurement, 40,* 755–759.

Mirkin, B. (2000, June). *Data driven classification and clustering.* Paper presented at the meeting of the Classification Society of North America, University of Montréal, Canada.

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20,* 359–363.

Morey, L. C., Blashfield, R. K., & Skinner, H. A. (1983). A comparison of cluster analysis techniques within a sequential validation framework. *Multivariate Behavioral Research, 18,* 309–329.

Murtagh, F. (2000). *Classification literature automated search service.* Available at http://www.pitt.edu/~csna.

Ng, R. T., & Han, J. (1994). Efficient and effective clustering methods for spatial data mining. In J. B. Bocca, M. Jarke, & C. Zaniolo (Eds.), *Proceedings of the 20th International Conference on Very Large Data Bases* (pp. 144–155). San Francisco: Morgan Kaufmann.

Okada, A. (1996). A review of cluster analysis research in Japan. In P. Arabie, L. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 271–294). River Edge, NJ: World Scientific Press.

Podani, J. (1989). New combinatorial clustering methods. *Vegetatio, 81,* 61–77.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika, 47,* 3–24.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20,* 134–148.

Rasmussen, E. (1992). Clustering algorithms. In W. B. Frakes & R. Baeza-Yates (Eds.), *Information retrieval: Data structures and algorithms.* Upper Saddle River, NJ: Prentice-Hall.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology, 53,* 94–101.

Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery, 1,* 317–327.

Sattath, S., & Tverksy, A. (1977). Additive similarity trees. *Psychometrika, 42,* 319–345.

Scheibler, D., & Schneider, W. (1985). Monte Carlo tests of the accuracy of cluster analysis algorithms: A comparison of hierarchical and nonhierarchical methods. *Multivariate Behavioral Research, 20,* 283–304.

Shepard, R. N. (1963). Analysis of proximities as a technique for the study of information processing in man. *Human Factors, 5,* 33–48.

Skinner, H. A. (1978). Differentiating the contribution of elevation, scatter, and shape in profile similarity. *Educational and Psychological Measurement, 38,* 297–308.

Sneath, P. H. A. (1980). The risk of not recognizing from ordinations that clusters are distinct. *Classification Society Bulletin, 4,* 22–43.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy.* San Francisco: W. F. Freeman.

Späth, M. (1980). *Cluster analysis algorithms for data reduction and classification of objects.* Chichester, UK: Ellis Horwood.

S-Plus. (1999). *S-Plus 2000 modern statistics and advanced graphics.* Seattle, WA: MathSoft Inc.

Storms, G., Van Mechelen, I., & De Boeck, P. (1994). Structural-analysis of the intension and extension of semantic concepts. *European Journal of Cognitive Psychology, 6,* 43–75.

Swanson, H. L. (1995). Effects of dynamic testing on the classification of learning-disabilities: The predictive and

discriminant validity of the Swanson cognitive processing test (S-Cpt). *Journal of Psychoeducational Assessment, 13,* 204–229.

Viswesvaran, M., Schmidt, F. L., & Deshpande, S. P. (1994). A metaanalytic method for testing hypotheses about clusters of decision-makers. *Organizational Behavior and Human Decision Processes, 58,* 304–321.

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

Wasserman, S., & Faust, K. (1994). *Social networks analysis: Methods and applications.* New York: Cambridge University Press.

Weinstein, J. N., Myers, T. G., O'Connor, P. M., Friend, S. H., Fornace, A. J., Kohn, K. W., Fojo, T., Bates, S. E., Rubinstein, L.V., Anderson, N. L., Buolamwini, J. K., van Osdol, W. W., Monks, A. P., Scudiero, D. A., Sausville, E. A., Zaharevitz, D. W., Bunow, B., Viswesvaran, V. N., Johnson, G. S., Wittes, R. E., & Paull, K. D. (1997). An information-intensive approach to the molecular pharmacology of cancer. *Science, 275,* 343–349.

Welch, W. J. (1982). Algorithmic complexity: Three NP-hard problems in computational statistics. *Journal of Statistical Computation and Simulation, 15,* 17–25.

Wolfe, J. H. (1970). Pattern clustering by multivariate mixture analysis. *Multivariate Behavioral Research, 5,* 329–350.

Wong, M. A. (1982). A hybrid clustering method for identifying high-density clusters. *Journal of the American Statistical Association, 77,* 841–847.

Wong, M. A., & Lane, T. (1983). A *k*th nearest neighbour clustering procedure. *Journal of the Royal Statistical Society, Series B, 45,* 362–368.

Zhang, T., Ramakrishnan, R., & Linvy, M. (1996). BIRCH: An efficient data clustering method for very large databases. In *Proceedings of the ACM SIGMOD Conference on the Management of Data* (pp. 103–104). New York: Association for Computing Machinery.