A

"How can I know if it's 'real'?" A Catalog of Diagnostics for Use with Three-Mode Factor Analysis and Multidimensional Scaling

Richard A. Harshman

Far too often, solutions obtained by multivariate procedures—including factor analysis, multidimensional scaling, and cluster analysis—are interpreted, and even published, without adequate evaluation of their reliability or validity. Particularly among inexperienced users, there is an uncritical and somewhat cavalier approach to determining what parts (or which version) of an analysis to accept. Clusters or dimensions are frequently taken to be "real" whenever an interpretation can be projected onto them by the imagination of the analyst. On the other hand, dimensions that don't fit preconceptions and are hard to interpret tend to be dismissed too easily. While some users may make a feeble attempt at justifying their choice of dimensionality by examining improvements in fit values, little effort is otherwise expended in determining whether clusters or dimensions are stable or reliable, whether the model is appropriate for the data, whether the algorithm achieved correct convergence, whether serious outliers are present in the data, and so forth.

On the other hand, more experienced and sophisticated users often do employ diagnostic checks, but the particular ones that are applied will differ from one user to the other. Many of the techniques employed by a given analyst may have been developed by the analyst himself, handed down by word-of-mouth, or picked up from a passing reference in a published article. Consequently, inexperienced users have little chance of being exposed to such procedures, except by apprenticing to someone who knows them or by reinventing the techniques themselves. While pub-

---

Major portions of this article were written at Bell Laboratories, Murray Hill, New Jersey. The author is grateful for Bell Laboratories' continuing support of research on three-way models and related topics.

lished accounts exist for some of these techniques, they are unfamiliar to most potential users; as yet, there is no common body of accepted, well-known methods available. And, because many editors are also unaware of the importance and proper use of diagnostics, they permit publication of articles that lack essential evidence for the validity of the solutions presented.

This is, to say the least, an unfortunate state of affairs, and this informal article is one attempt to initiate changes. In the following discussion, I will first point out the role that diagnostics can and should play and the questions that they can answer. I will then briefly try to convey the range and nature of diagnostic procedures that can be used in factor analysis and MDS by providing an informal list and brief description of those procedures of which I am aware, including brief mention of some still being developed. As we shall see, a number of different methods can be used to check the optimality, reliability, and validity of a three-mode analysis solution. In fact, the potential variety of diagnostic techniques for cluster analysis, three-mode factor analysis and MDS is so substantial that it could (and hopefully will) constitute a major area of growth and refinement of multivariate methodology in the next few years.

In this informal listing or "catalog," an attempt is made to develop a natural classification scheme for the diagnostics, based on the type of information they need (such as the data itself, the factor loadings resulting from a single analysis, loadings from several analyses, and so on) and which aspects of the solution they focus on (including the fitted parameters of the model, the residuals from the fit, or indices of overall goodness-of-fit). The catalog is intended only as an introduction or overview of some of the possible diagnostics that can be used at various stages of an analysis. It certainly is not a thorough exposition; detailed questions of how to use these diagnostics are not covered, although brief descriptions of applications are sometimes given.

The objective is to increase awareness of the methods, to generate interest and discussion, and to suggest methods that have been omitted from the list—perhaps as an initial step toward a more complete treatment in the future—and to encourage investigators to begin using these diagnostics. For some of the procedures, such as diagnostic interpretations of loadings patterns, enough information is provided so that investigators should be able to use the methods to guide them during three-way PARAFAC analyses. For other more esoteric or incompletely developed procedures, only enough information is provided to suggest the possibilities that may later become perfected.

Diagnostics for cluster analysis are not considered here. While proper diagnostics are just as important in cluster analysis as they are in factor analysis and multidimensional scaling, cluster-analytic diagnostics often involve somewhat different procedures. In order to avoid undue complication, this article focuses on the use of diagnostic procedures in factor analysis and multidimensional scaling, especially those that are well suited for the three-way intrinsic axis methods PARAFAC-CANDECOMP and INDSCAL. Many of the techniques discussed, however, would also be suitable for Tucker's model and for other three- and two-way

methods. The appropriate adaptations should usually be apparent to the reader.

OBJECTIVES FOR DIAGNOSTIC PROCEDURES

Two Roles for Diagnostics

Diagnostics have both an *exploratory* and a *confirmatory* role. In their exploratory role, they provide guidance during the conduct of an analysis. With such guidance, one can adjust the parameters of successive analyses to maximize sensitivity to the characteristics of the data, rather than stop with the automatic application of standard data preprocessing and data-analytic options. In their confirmatory role, they lend greater assurance to any conclusions drawn from the final factor-analytic or MDS solution.

Exploratory Role

Most of the diagnostics to be discussed in this article are not merely directed at establishing the reliability and validity of a solution that has already been selected for interpretation. Rather, they are important tools for deciding "what to do next" as one proceeds through the course of the analysis. This use of diagnostics goes hand-in-hand with an *interactive* view of data analysis. From this perspective, a factor-analytic or MDS analysis is not a "one-shot" application of an automatic procedure but a multiple-stage interplay between the data analyst and the data, involving repeated cycling between application of the program for analysis and application of diagnostics for evaluating the outcome of the latest stage of the analysis. In this role, diagnostics answer questions such as: Are more iterations required? Should one extract additional dimensions? Are orthogonality constraints necessary?

Confirmatory Role

The confirmatory use of diagnostics for factor analysis and MDS is important to protect both the user and the scientific community from misleading and inappropriate solutions. When used in this way, diagnostics confirm that the analysis model appears appropriate for the data and that the resulting solution is reliable, optimal, and generalizable. Hopefully, as we become more sophisticated, such confirmation will come to be viewed as an essential part of any solution, and future editors will expect a written description of the diagnostics used and the results obtained, as part of the necessary supporting evidence for the conclusions drawn from any analysis. Thus, as familiarity with diagnostics grows, minimal standards of evidence for reliability and validity should develop, helping to screen out some of the most meaningless applications of factor analysis and MDS techniques from the literature.

General Objectives

When using diagnostics, there are at least four basic things that one seeks to determine about a given analysis: (a) theoretical appropriateness, (b) computational correctness, (c) statistical reliability, and (d) explanatory validity. It is important to establish these characteristics for any analysis, be it two- or three-way, dimensional or cluster. However, here we formulate the problem only in terms of three-way factor analysis and multidimensional scaling, particularly the intrinsic axis methods PARAFAC-CANDECOMP and INDSCAL. To elucidate each of the four basic characteristics listed above, we discuss it in the context of specific questions concerning a three-way analysis problem.

Questions of Theoretical Appropriateness

1. How appropriate is our basic multivariate model?

   a. Is the analysis model we intend to employ appropriate for the kind of question we want to ask? For example, are we really looking for latent dimensions, or would taxonomic clusters be more appropriate? (If clusters, would we want hierarchical or additive, disjoint or overlapping?) Do we want additive main effects and unrestricted interactions, such as an ANOVA would provide, or the structured kind of multiplicative interactions that multidimensional models provide, or both?

      This one set of questions must be answered without diagnostics, before any analysis has been started.

   b. Is the analysis model we seek to employ appropriate for the data? Is the structure underlying the observed relationships more treelike, implying a cluster model, or more spacelike, implying a factor-analytic or MDS model.

      This question is still largely theoretical, but some diagnostics are beginning to be developed to aid the investigator in making this decision.

2. Which of the various three-mode factor-analytic or MDS models is most appropriate for these data?

   a. Is our data distancelike, so MDS is appropriate, or profilelike so that factor analysis is appropriate, or is it scalar-product or covariancelike, again calling for factor-analytic models?

   b. Are the three-mode data likely to provide the appropriate pattern of variations in latent factors across all three modes, so that unique determination of axes by PARAFAC or INDSCAL is possible? Or, instead, is the third mode simply a set of replications differing only by

random variations? Or is some intermediate, partial, or more complex variation pattern likely?

c. If there are genuine differences between the levels of the third mode, are these likely to be appropriate for the strong "system variation" model that allows direct fitting, or is the "object variation" model more appropriate, calling for indirect fitting?

d. Would Tucker's three-mode T2 or T3 model be more appropriate for these data? Would complex variations in factor obliqueness or factor interaction contribute a major part of the data variance?

e. If distancelike data are involved, should these be considered squared distances?

f. If the rows and columns correspond to the same set of entities, are the relationships among them symmetrical (that is, does $x_{ij} = x_{ji}$?), or are there systematic asymmetries that need to be described, perhaps calling for a more general model such as DEDICOM?

g. Should one consider a model that specifies orthogonal loadings in one or more modes—for instance, to fit a hierarchical factor solution?

3. Should the data be preprocessed or reexpressed in particular ways to make it appropriate for the model or to bring out its most interesting properties?

   a. Is the data likely to contain conditional origins, additive constants, and two-way interactions requiring centering of one or more modes to make it appropriate for the ratio-scale model?

   b. Should the variances or mean-squares of the variables, subjects, occasions, or whatever be standardized?

   c. Is reweighting of variables, subjects, or other subsets of the data desirable to minimize the influence of unreliable data or to stress aspects of the data where good fit is most important?

   d. Should nonlinear transformations be employed, such as log transformations?

Questions of Computational Correctness

4. Has a particular iterative fitting procedure converged to the desired optimum?

   a. Are there indications of incomplete convergence due to an overly lax convergence criterion?

b. Is there very slow convergence of some solutions because of a bad starting position or of all solutions because of certain properties of the data?

c. Once convergence has been established, is it convergence to an uninteresting local optimum or to one of several competing interesting solutions? Or is there well-behaved convergence to the same solution regardless of starting position?

5. To what extent is the solution independent of starting position? Do some parameter values change as a function of starting position, while others are fairly stable? Is the configuration of points after rotation to congruence independent of starting position? If we are using intrinsic axis methods, do we obtain the same orientation of axes regardless of starting position of the iterative procedure, or is the orientation dependent on starting position (and, perhaps, not uniquely determined by the data) for some or all dimensions?

Questions of Statistical Reliability

6. What is the stability of the solution across subsamples of the data? If some characteristics are more stable than others, which details of the solution are stable enough to justify interpretation? Which conclusions are generalizable to new sets of subjects, or variables, or occasions, and so on? What kinds of generalizability do we have (for instance, across subjects, across variables, across occasions, and the like); which kinds do we desire?

7. Have we chosen the correct dimensionality? Are all our dimensions stable enough to be recognized in two split-half subsamples? If not, are there other compelling reasons to retain any of the unreliable ones? Alternatively, are there further stable dimensions that we are overlooking?

Questions of Explanatory Validity

8. Are the results interpretable? How are the obtained dimensions related to outside information about the variables, stimuli, individuals, and so on? Are our problems of interpretation, if any, more likely due to the data, to the analysis procedure, or to limitations of our understanding?

9. Are there nonlinear relationships among dimensions that would indicate the appropriateness of a nonlinear model of lower dimensionality? Should we consider nonlinear reexpression of the data?

10. What are the properties of the residuals?

    a. Do they have patterns indicating data structure not captured by the current analysis? Should ANOVA, cluster analysis, or some other procedure be applied to the residuals in order to uncover features that the factor analysis would miss?

    b. Are there extreme outliers (or groups of outliers) in the data that might have unduly biased the solution? Should some data points be replaced by missing-data codes and the analysis run again?

11. Do properties of the obtained loadings indicate that a different version of the three-way model—such as a different extended model—would have been more appropriate? Should different data preprocessing methods have been used?

12. If the current analysis has produced a reliable, meaningful solution, what new experiments could be conducted to test the hypotheses emerging from these results? How could other available data confirm or conflict with the general conclusions you have drawn?

OUTLINE OF THE CATALOG OF DIAGNOSTIC TECHNIQUES

There is a large variety of techniques that can be used to help answer the questions listed above. Although there is not space in this appendix to provide a detailed explanation of all these techniques, I will provide an informal list or catalog of those known to me and will sometimes include a brief statement of how they are used. But before presenting the catalog, I will first give an overview, in outline form, as a guide to the more detailed discussion to follow.

The catalog is organized into four subsections, based on the type of information being examined by the diagnostic procedures in that section:

I. *Zero-fit Diagnostics:* Data diagnostics to be performed before doing the three-way analysis.
II. *One-fit Diagnostics:* Analysis diagnostics based on examination of the results of a single fit (such as an analysis from a particular starting position, at a particular dimensionality, using a particular set of analysis options).
III. *Many-fit (single data set) Diagnostics:* Analysis diagnostics based on comparisons across several different fits made to the same data set (such as using different random starting positions, different dimensionalities, or different analysis options).
IV. *Many-fit (many data sets) Diagnostics:* Analysis diagnostics based on comparisons across fits made to different or partially different data sets (such as random split-halves of the subject sample, or stimulus sample, or overlapping subsamples such as used in jackknifing and bootstrapping).

For the analysis diagnostics in categories II and III, the procedures are further categorized according to whether they are based on examination of the loadings, examination of the residuals, or examination of the overall goodness-of-fit measures. For section IV, they are broken down into several different kinds of reliability evaluation procedures. This system of classification might be helpful for finding the proper diagnostic to use when one is considering a particular part of the computer output or for highlighting certain logical or mathematical relationships among procedures. The outline of the catalog is as follows.

I. *Zero-fit Diagnostics*
   - A. Checking for outliers
   - B. Checking reliability of the data
   - C. Determining the structural form of the data
   - D. If data are distancelike, checking tree versus spatial models
   - E. Checking appropriateness of three-way models using two-way "collapsed" versions of the data
     1. Comparison of the dimensionality of different two-way "collapsed" versions of the data—are the estimates consistent?
     2. Comparison of the configurations (that is, dimensions after rotation to maximum agreement) across different two-way "collapsed" versions of the data —do special two-way interaction dimensions emerge in certain pairs of modes?
     3. Comparison of the dimensions (after rotation to agreement) found in different slices of a given mode.

II. *One-fit Diagnostics*
   - A. Based on examination of loadings
     1. High correlation among factors
     2. Constant factors
     3. Nonlinear relationships among factors
     4. Interpretability
     5. Convergence checking
   - B. Based on examination of residuals
     1. Examining patterns in the relative sizes of mean square error (MSE) for different levels of each mode
     2. Outputting residuals for detailed study
   - C. Based on overall goodness-of-fit values
     1. Comparison of R with estimated data reliability
     2. Comparison of the square root of MSE with expected size of errors
     3. Comparison of R and MSE

III. *Many-fit (single data set) Diagnostics*
   - A. Based on comparison of loadings
     1. Across successive iterations
     2. Across different solutions (different random starts) at a given dimensionality
     3. Across different dimensionalities (to study evolution

of dimensional structure and interpretations)
- B. Based on comparison of residuals
  1. Across competing solutions at a given dimensionality
  2. Across dimensionalities (to see for which part of the data the fit improves when adding a particular dimension)
  3. In terms of error distributions
- C. Based on comparison of overall fit values
  1. Across iterations (to check convergence)
  2. Across solutions—identifying local optima, incomplete convergence, and so on
  3. Across dimensionalities—classic search for "elbow" in fit versus dimensionality curve

IV. *Many-fit (many data sets) Diagnostics*
- A. Comparisons across split-halves of the data
  1. When to split
  2. How to split
  3. How to compare across splits
- B. Resampling methods of estimating reliability: bootstrapping and jackknifing
  1. Jackknifing
  2. Bootstrapping
- C. Cross-validation techniques testing fit when dimensions are applied to a new sample
- D. Randomization tests—comparison of "shuffled" data with observed data to obtain significance tests for three-way variation
- E. Comparison of analyses across experiments

It is hoped that this system of classification will be generally helpful and may loosely correspond to the order in which some of the tests may be conducted. However, it is not intended to provide an actual strategy for the interactive use of these diagnostics or for guiding one's choices while performing an analysis. Such strategy questions are not considered in detail in this paper.

CATALOG OF DIAGNOSTIC TECHNIQUES

In this section, an attempt has been made to list the various diagnostic comparisons, tests, or procedures that can be employed to help answer the questions presented in the previous section.

I. *Zero-fit or Data Diagnostics:* To be performed before doing the three-mode analysis.
- A. Examination of data to detect outliers, with the possibility of omitting outliers from the analysis by declaring them as missing data (while saving their identity for subsequent study).
- B. Evaluation of reliability of the data—by comparing replications, computing test-retest correlations for each subject or variable, or by comparison of corresponding

cells on either side of the diagonal of supposedly symmetrical data, and so on. Possible elimination of subjects, conditions, and the like that have insufficient reliability. Also use reliability estimates for comparison to later variance accounted for by the model.

C. Determination (mostly on the basis of outside information or theory) of whether the data is likely to be more distancelike, profilelike, or scalar-productlike, so that the appropriate analysis model and procedure can be selected.

D. If the data are thought to be distancelike, evaluating whether a tree-structure or spatial model for the distances is more appropriate; this can be approached by fitting the alternative models and comparing fit, but recent work of Carroll, Pruzansky, Tversky and others is developing some test statistics that can be computed from the data itself, such as the skewness of the distribution of distances.

E. Examination of systematicity of variation across each mode, by means of two-way analysis of "collapsed" versions of the three-way data; to help determine which kinds of three-way variation are present and thus which three-way model is useful. This approach is still largely untested, and these procedures are not as essential as others to be discussed later, but the following are examples of steps that might be useful:

1. Plots of successive eigenvalues or singular values of data collapsed across various modes; do the plots reveal systematic dimensionality in all three modes? Is dimensionality similar across different modes, or might Tucker's model be more appropriate?
2. Examination (and possibly canonical correlation) of eigenvectors or singular vectors extracted from various two-way "strung out" or "collapsed" versions of the data; how are these related? For example, do the Mode *A* vectors of the *A-B* collapsed data resemble the Mode *A* vectors of the *A-C* collapsed data?
3. Canonical correlation of eigenvectors or singular vectors extracted from successive layers of the data sliced across a given mode; how similar are the dimensions in the different layers of the three-way array?

II. *One-fit Diagnostics:* Analysis diagnostics that utilize results of a single fit—for example, diagnostics for a solution obtained from a particular starting position, at a particular dimensionality, using a particular set of analysis options.

A. Examination of the fitted parameters—the factor loadings.*

1. Examination of correlations and cross-products among factor loadings for each mode.

---

*Points 1–3 are specific to an intrinsic axis model.

a) Check for very high correlations between two (or more) dimensions in all three modes ("very high" means, roughly, above .8).
   (1) If the pattern is consistently obtained across starting positions, and the triple product of the correlations for all three modes is negative (that is, either one correlation is negative or all three are negative), this could signal a degenerate solution (as discussed in chapter 6). If the data have not been centered on one or two modes, try additional centering. If careful centering and standardization does not remove the degeneracy, then a "hard-core" degeneracy may be present, requiring orthogonality constraints on at least one mode.
   (2) If the pattern only emerges at a relatively high dimensionality and is not consistent across solutions or split-halves of the data, it might simply indicate that more factors are being extracted than can be supported by the data. In this case, the triple product of correlations should as often be positive as negative. If, after applying orthogonally constraints, the dimensions in question are still not interpretable and are not similar across solutions or split-halves, then this suspicion is confirmed. But if orthogonally constraining one mode gives similar results across split-halves, then dimensionality is not too high (see example in appendix C).

b) Check for high correlations (above .6 or so) between all or most dimensions, in all three modes. This can indicate that too many dimensions have been extracted and/or a general lack of uniqueness of the solution, resulting in arbitrary (correlated) combinations of the "true" underlying dimensions appearing in all three modes. If the triple product of correlations is often positive and different patterns of correlations are observed from different starting positions and across split-halves, this is probably not a "degenerate" solution in the restricted sense used by Harshman and Lundy (chapter 6). Check if all dimensions of orthogonally constrained solutions are replicable across split-halves; if not, reduce dimensionality. (However, if the configuration is reliable but the axis orientation is not, as indicated by high canonical correlations across split-halves, the problem is not extraction of too many dimensions but rather lack of conditions producing unique axes. You must use additional levels of

data or some external rotation criteria to obtain unique axis orientation.)

c) Check for high correlations between particular dimensions in one mode only, with lower correlations between the same dimensions in the other two modes. This does not indicate "degeneracy" or that too many dimensions have been extracted. Rather, it warns that certain factors may not show the distinct patterns of variation necessary to determine axis orientation uniquely. (Note: Even if the correlated dimensions are nonunique, the rest of the solution could be uniquely determined.)

2. Examination of loadings, to check for one or more "constant" factors (that is, factors for which all the loadings are approximately the same size).

a) A factor that is constant—that is, it has loadings of constant size and sign—in all three modes indicates an overall additive constant in the data that probably should be removed. The size of this additive constant can sometimes be estimated from the triple product across modes of the factor's loadings. However, it is safer to simply apply centering to one or more modes.

b) A factor that is constant in two modes, with varying loadings in a third mode, indicates a different additive constant for each slice (when sliced so that the varying loadings correspond to different slices of the data). Centering across one or more of the constant modes should remove this factor.

c) A factor that is constant in only one mode simply indicates a "true" factor that doesn't vary much across that mode. If there is only one factor that is constant in any given mode, then such a factor may be uniquely determined; however, comparison across starting positions and split-halves would be desirable to confirm stability. If two (or more) factors have nearly constant loadings within the same mode, then their loading patterns may not be uniquely resolved in the other two modes, and, sometimes, other nonconstant factors will be "contaminated" as well. Centering across the mode in which the constant loading patterns occur will eliminate these factors from the data and may improve the recovery of the other factors. (Note: A more detailed discussion of loading patterns and their correspondence to main effects, 2-way interactions, and 3-way interactions, along with a discussion of the effects of various centering schemes on such factors, appears in chapter 6.)

3. Comparison of loadings across factors to check for nonlinear relationships.

Such relationships indicate either that the data needs to be transformed nonlinearly and/or that the actual latent factors combine nonlinearly to create the observed data. Particular conclusions, including the form of likely nonlinear relationships, can be deduced from the study of the plots of loadings of one factor against another or against functions of several other factors. Nonlinear factor models can sometimes be constructed on the basis of this type of information.

4. Preliminary checking of interpretability of dimensions for the given starting position, by detection of meaningful and/or expected patterns of loadings in each mode.

   This can provide support for (or cast doubt upon) the validity of a particular solution. This can also provide insight into relationships among different solutions (such as why certain dimensions split into more specific ones as dimensionality is increased, or which aspects of a solution replicate across split-halves, or whether alternative meaningful solutions appear in different competing—locally optimal—solutions). However, an extensive study of interpretability should normally be deferred until other diagnostics indicate that one has a candidate for an optimal solution (see below).

5. Checking convergence by examining the rate at which loadings are changing across successive iterations.

   While this is technically a multiple-fit comparison, it is accomplished within an analysis from a given starting point, and so will be briefly mentioned here. Convergence rates can vary greatly, depending on the particular data being analyzed and sometimes on starting position for a given analysis. If the factor loadings are continuing to change in small steady steps that do not diminish appreciably in size (for instance, steps of 1% change per iteration), then one cannot be confident that the given solution is essentially the same as one that would be obtained after many more iterations. Comparison of a given set of loadings with the loadings obtained 10, 20 or even 50 iterations previously should indicate if the loadings are "settling down" or if they are continuing to slowly drift. More accurate evaluation of convergence can be made by use of multiple starting points (see below).

B. Examination of residuals.

   1. The Mean Squared Error (MSE) for the solution should be of "reasonable" size, based on expected precision of measurement, likely reliability of the data, and so on.
   2. An "Error Analysis Table," which prints out mean-squared error for each level of each mode (for

instance, for all the data points that involve a particular stimulus or all the data points for a particular subject, and so on) can be used to see whether there are certain portions of the data that are causing particular problems. In examining this table, the following points should be checked:

a) Are any MSE values very high or very low, relative to the others?
b) Are there systematic patterns in the larger versus smaller MSE values? Do certain aspects of the data cause problems, including unreliable groups of subjects, certain "difficult" types of stimuli, and the like?
c) If the data being analyzed have unequal variances or mean-squares across the levels of a given mode, the MSE differences among the levels of that mode can be expected to mirror these input differences. Levels that have higher variance or mean square on input will usually have higher MSE on output. Thus, comparison of MSE values should take the input mean-squares or variances into account. One way to do this is to look at Stress at each level rather than "raw" MSE. This problem does not arise, of course, if the data have been size-standardized within each level of the mode in question.
d) If, in each mode, one or two very high MSE values stand out, then the intersections of these levels should be examined for outliers far out of range (such as keypunching errors). For example, if the MSEs for variable 7, person 2, and occasion 2 are all very high, then the data point for person 2 measured on variable 7 on occasion 2 should be examined. (Note: Some programs—including PARAFAC—contain an option to check each data point on input against a user-specified range of possible valid values. Points outside this range will be identified in the output and treated as missing values during the analysis. It is recommended that such an option be used whenever possible—for instance, whenever the data consists of questionnaire responses using a fixed response scale. The points identified as outside the valid range should then be checked and corrected.)

3. For more detailed analysis of residuals, some programs (including PARAFAC) provide the option to write out the entire set of residuals on an output disk file of the user's choice. When this is done, the residuals can be examined in detail by using a variety of standard analysis and graphics programs. Residuals can be plotted against original data values, a histogram of their distribution can be plotted, and so on. Also, the residuals so output

can subsequently be input for further analysis by more sophisticated procedures (such as cluster analysis) or can even be used as input for another round of three-way analysis, which would extract a set of additional factors with fitted ($\hat{x}$) values completely orthogonal to the first set.

C. Examination of overall goodness-of-fit indices—for instance, correlation between the data and the predicted data, Stress, the ratio of the MSE to the mean-square data value, and so forth.
   1. How does the *R* value compare to the test-retest reliability of the data (either known or conjectured)? When sufficient dimensions are extracted, it should be approximately the same as the expected reliability of the data being analyzed, if the model provides a good description of the systematic part of the data. Keep in mind, however, that interesting and informative structure can sometimes be recovered even when much of the systematic part of the data cannot be explained. Thus, an interpretable solution should not be rejected outright simply because the *R* was, for example, less than half of the data reliability.
   2. How does the square root of the MSE value (that is, standard deviation of the error) compare to the expected size of typical error values? For example, root-MSE of 1 to 2 might be reasonable for a 9-point scale.
   3. How do *R* and MSE compare? Usually, when *R* is high, MSE will be small. However, when there are a small number of very errant data points, it is possible to get high *R* values (fitting the very large error variance due to these errant points) but also very large and undesirable MSE values. Both should be checked for reasonableness. (Comparing variations of *R* and MSE values across subsets of the data provides a particularly useful check; see below.)

III. *Many-fit (single data set) Diagnostics:* Analysis diagnostics based on comparison across multiple fits made to the same data set—for example, using different random starting positions, different dimensionalities, different analysis options, and so on.

A. By examination of the loadings.
   1. Comparison of different stages in the iterative process to assess convergence (this was also mentioned in II.A.5 above).
      a) For assessing convergence, loading changes are more important than changes in fit values. Fit values will rapidly improve in the early iterations of an analysis and then level off and show only gradual improvement in later iterations (when correct rotation is being established). These gradual changes in fit values can occur

when some of the loadings themselves are still changing substantially. But since it is the loadings that one will be interpreting, these are the quantities that must be stabilized in order to consider a solution to be properly converged.

b) Loading changes should be small at "convergence"; one standard option is to compare changes due to one iteration with the RMS average loading size for each dimension. A conservative rule of thumb is that no loading should change more than .1% of the RMS average size of a loading on that dimension (in that mode) from one iteration to the next; alternatively, the change should not exceed 1% across 10 iterations. This rule can be relaxed to allow 5% or more, with well-behaved, quick converging data, but with such data, relaxing the criterion is often unnecessary because convergence is rapid anyway. Unfortunately, it is with slow converging data that a more stringent criterion is sometimes necessary to prevent a misleading premature fulfillment of the convergence test. (Premature declarations of "convergence" can be detected by comparison of results obtained from different random starts; see III.2, below.)

c) Certain patterns of change can indicate convergence difficulties in particular subsets of dimensions ("subspaces" of the solution). It often happens that loadings for certain factors converge more quickly than others. However, it may occasionally occur that a few factors almost never converge, although the rest of the solution is stable. In this case, different starting positions and/or different dimensionalities should be tried. Are these "difficult" dimensions highly correlated in a single mode, indicating insufficient independent variation to provide uniqueness of axes, or are they highly correlated in all three modes, suggesting either degeneracies (if the triple product of the correlations is persistently negative) or extraction of too many factors (otherwise)?

2. Comparison of several allegedly converged solutions obtained from different random starting places to confirm convergence. Solutions that change slowly enough to meet the convergence criterion may have converged and come close to their terminal values. Alternatively, they may have become trapped in a difficult "ridge" on the surface that they are trying to climb and consequently may be experiencing slow convergence at values very different from the ones they would eventually take on if allowed to iterate indefinitely. Comparison of results obtained from

different random starting positions can help in discriminating these two situations and also provide information on the strength of determination of unique axes for different dimensions.

a) Do several solutions (3–6) agree "closely enough" for their differences to have no effect on interpretation? If so, convergence and uniqueness are both indicated.

b) If several solutions (3–6) agree approximately—that is, they are "going toward the same place"—then setting a more strict convergence criterion and continuing to iterate on any one of the solutions should provide an accurate estimate of the more fully converged solution that would have been obtained with all of them.

c) Do the solutions fall in two or more groups? For example, did two random starts give one solution and four random starts give another? If so, which dimensions are the same in both groups; how do the other dimensions differ? The presence of competing solutions found from different random starting positions could have several interpretations. Check the following possibilities:

   (1) One or more of the groups might represent a local optimum where the program repeatedly "got stuck" on its iterative upward search for the globally optimum set of loadings. Compare the fit values for the competing solutions. Are there differences in the second decimal place? If so, the set or sets with the lower fit value may be a local optimum.

   (2) How do the interpretations of each of the two or more competing solutions relate to those of solutions obtained at lower and higher dimensionalities? Are the dimensions in both competing sets interpretable? Do they represent two different subsets of a larger common set of dimensions that will be obtained in higher dimensional solutions? (Obviously, one must wait till more dimensions are extracted to check this.) If so, both competing solutions may represent "valid" but incomplete approximations to the higher dimensional "true" solution. (Note, however, that the form of the dimensions will often be clearer in the higher dimensional solution, where all valid dimensions emerge, since in such a solution no dimensions need be distorted to help adjust for the effects of a dimension not yet extracted. On the other hand, if determination of axis orientation is weak at the higher dimensionality, the dimensions

in this solution may seem less clear. When too many dimensions are extracted, the form of some of the dimensions often starts to break down.)

(3) Are there two (or more) competing solutions, with one (or more) of them showing very high factor intercorrelations or other signs of break down? The well-behaved and interpretable solution should be preferred, particularly if it has a higher fit value. (Occasionally, the well-behaved solution may have a slightly poorer fit value; nonetheless, it should probably still be preferred.)

(4) Are there many different solutions (almost as many as random starting positions)? Across starting positions, do certain sets of factors keep changing—recombining differently? This could indicate that those particular factors do not have a unique rotation determined by the data.

(5) Are all or almost all factors changing across different starting positions? This would indicate rotational indeterminacy of the solution as a whole and suggests either that one mode of the data does not have the required systematic variation of any factors necessary to establish uniqueness, or, alternatively, that substantially more dimensions are being extracted than can be supported by the data (for instance, at least 50% too many dimensions).

3. Systematic comparison of dimensions across solutions of different dimensionalities can shed light on "family" relationships among dimensions in the higher dimensional solution and reveal stronger versus weaker aspects of the solution.

a) It may sometimes be useful to construct a "Terbeek tree," showing which of the dimensions in the two-dimensional solution (for example, were present in the three-dimensional solution) and so on. Display the correlations or other similarity measures between the dimensions at several different dimensionalities, from one up to the maximum number extracted. Such a "tree" of dimensional relationships will reveal at each level whether a dimension split into two dimensions, whether components of several dimensions were drawn off to form a new dimension, or whether an entirely new dimension emerged. In this way, the tree can relate the dimensions at all different dimensionalities. Hopefully, this will help one to understand the process by which various dimensions emerge. It may also suggest that at a

certain dimensionality, a particular dimension is "contaminated" by other specific unextracted dimensions or is otherwise distorted.

b) Is some sort of meaningful hierarchical structure suggested by the tree? How related are the *interpretations* of two dimensions that "emerged" from a common ancestor in a lower dimensional solution? Can these "family" relationships among dimensions themselves suggest interesting interpretations of the data, much as a hierarchical cluster analysis would?

c) At what dimensionality was the interpretation of a given dimension clearest?

B. By examination of the residuals.

1. Comparison of residuals (or MSE values for specific levels of each mode) across competing solutions at a given dimensionality, to determine which parts of the data are being fit by each solution. This may occasionally be useful when one of two competing solutions is suspected to be due to peculiar characteristics in a very restricted part of the data. If, in one of the competing solutions, the fit value is clearly lower for this part of the data but not for others, the suspicion is supported.
2. Comparison of residuals (or MSE values for specific levels of each mode) across dimensionalities to see which part of the data has reduced fit when each dimension is added. This will sometimes reveal that one of the smaller dimensions extracted at higher dimensionalities is mainly accounting for the variance of a single subject, variable, or whatever. It might also show that the dimension is attempting to account for certain "outlier" points for several subjects.
3. Comparison of distributions of the residuals at different dimensionalities. Check, in particular, whether there are still many large values in the tails of the distribution at lower dimensionalities; hopefully, these become less common as more dimensions are extracted, approaching the desired low frequency compatible with normally distributed error at the "correct" dimensionality (unless other kinds of structure are present that cannot be fit by the factor model).
4. For more methods of looking at residuals of a PARAFAC analysis, see Kettenring's article (cited in chapter 5).

C. By examination of overall fit values (*R*, *R*-squared, Stress, MSE, and so forth).

1. Comparison of fit values across successive iterations to determine convergence. This is the "traditional" method of assessing convergence of the solution. As noted earlier, however, this method is not recommended for our three-way intrinsic axis models, since the fit values often change quite

slowly during the later stage of a given analysis, while the loadings themselves are still undergoing considerable modification (due often to shifts of axis orientations). However, in certain degenerate solutions, the loadings of highly correlated factors will continue to change, even though the improvement in the fit is negligible (such as in the seventh decimal place). In these cases, it is pointless to wait for "convergence," since mathematical analysis shows that there may be no local optimum. Instead, compare solutions from several starting places to determine that after two or three hundred iterations they show similar "nonconverged" loading patterns.

2. Comparison of fit values across solutions obtained at the same dimensionality but from different starting positions
   a) This permits evaluation of the relative progress of several different solutions toward a common solution; if all solutions in a given set are "going to the same place," then pick the one with the highest fit value for interpretation or for continuation of the analysis with additional iterations
   b) If there are two or more competing solutions at a given dimensionality, comparison of overall fit values may indicate that the one that is difficult to interpret is in fact an uninteresting local optimum with substantially lower fit. Note, however, that two solutions that differ in fit might both be interpretable or "interesting" local optima; they may select different subsets of dimensions from a larger set that will be revealed at a higher dimensionality.
3. Comparison of fit values across dimensionalities in an attempt to determine the best number of dimensions to extract from a given data set. This is the classic "scree" test or search for the "elbow" in the fit-versus-dimensionality curve.
   a) The scree test remains a very important method of assessing dimensionality. It has a straightforward logical rationale and a simple graphical method of implementation that make it easy to understand and apply: Improvements in fit (such as changes in R-squared) due to each additional dimension are plotted against dimensionality. When the points begin to fall onto a smooth line, one assumes that "real" dimensions are no longer being extracted; the small steady increments are presumably due to fitting error. (Only the points that deviate from the smooth fit-change-versus-dimensionality line should be taken to indicate "real" dimensions; the older approach of including the first point on the smooth "scree" line should

not be followed.) Monte Carlo studies indicate that the scree test, when used with care, is one of the most accurate methods of assessing dimensionality.

Because the scree test is based on fit values, it provides information that is complementary to that provided by tests that are based on replication of a pattern of factor loadings (including split-half, bootstrap, or jackknife methods). It will sometimes work when they fail, and vice versa. For example, when lack of independent variation in the three-mode data causes the orientation of some axes not to be well determined, then two split-half solutions may differ in rotation for some dimensions, and the "true" dimensionality may be underestimated if one stops extracting dimensions when loading patterns fail to replicate. However, the scree test does not depend on proper or consistent alignment of axes, since the fit values are still well determined even when axis orientation is not. Thus, the scree test can indicate the presence of a systematic configuration in higher dimensions, even when the axis orientations fail to replicate. In this situation, replication tests can only verify the higher-dimensional configuration if they are strengthened by inclusion of rotation-to-congruence procedures or if regression or canonical correlation is used to find comparable dimensions across solutions. But it is not altogether a weakness of the replication tests that they will not detect these higher dimensions, since axes that are unstable should not be interpreted. If one decides to apply some rotation procedure such as VARIMAX to determine axis orientation in situations in which the intrinsic axis property cannot be used, then replication of the VARIMAX-rotated axes provides an appropriate test of which loading patterns can be taken as sufficiently stable to interpret.

Although we have been considering the situation in which the scree test indicates a higher dimensionality than the replication tests, the opposite situation can also arise. A smooth elbow on the curve can make the scree point hard to identity. Dimensions that account for only a small portion of the variance may not show up clearly on the scree curve yet may be replicable and theoretically important (for instance, see Gandour and Harshman, cited in chapter 5). When such dimensions replicate, they should be included in the solution and interpreted.

In general, then, the scree and replication (e.g., split-half) tests of dimensionality are complementary and, whenever possible, should both be applied and the results compared.

b) Unpublished Monte Carlo tests of different fit measures suggest that while most fit measures gave similar results, *R*-squared or variance accounted for provides the clearest "elbow" at the true dimensionality of synthetic data and the flattest curve thereafter.

IV. *Many-fit (many data sets) Diagnostics:* Analysis diagnostics based on comparisons of fits made to different or partially different data sets—for instance, random split-halves of subjects or stimuli, overlapping subsets as used in bootstrapping or jackknifing, and so on.

This is a particularly important diagnostic technique. It gives the strongest basis for deciding what aspects of a solution are statistically reliable and thus potentially generalizable to new samples.

In the exploratory mode, these diagnostics can be used to help determine optimum characteristics of an analysis, such as dimensionality of a solution. When additional dimensions replicate in two split-halves, for example, then they deserve interpretation and probably should be included in the solution in some fashion. When dimensions do not replicate, they should only be interpreted with caution, and the investigator should seriously consider reducing the dimensionality of the solution.

In the confirmatory mode, these diagnostics assure the investigator that the obtained dimensions are not simply based on fitting noise in the data. Characteristics of the solution that are demonstrably stable across samples are in some sense "real"; they are characteristics of some larger population. Those that are not stable in a particular test may or may not be "real"; their apparent instability may be due to small sample size, and the variations across subsamples may be reduced by taking larger samples. On the other hand, the so-called "characteristics" may not show up consistently in new samples of any size, because they are due to random error. Thus, interpretations based on characteristics that do not replicate may not be justified; they may be attempts to interpret random sampling fluctuations.

It is worth noting that the most important form of stability that should be tested by these techniques is stability of *conclusions* or *interpretations*. The key objective of any method of evaluating reliability—be it split-half, bootstrapping, or whatever—is to determine whether the important points of interpretation (such as the scientific conclusions or the recommendations for action) that are drawn from a particular sample are justified, at least to the extent that they can be generalized to equivalent samples.

In addition to this qualitative evaluation of the "robustness" of conclusions, more quantitative methods of measur-

ing stability can often be used to place confidence bounds around the values of particular parameters in a given solution. These techniques are usually applied to assess the reliability of the loadings, although they can also be used to place confidence bounds around fit values, angles between dimensions, and the like. There are several different methods of assessing reliability of factor-analytic and MDS solutions, including analysis of split-halves, resampling methods such as bootstrapping and jackknifing, cross-validation by application of a given set of loadings to a new data set, randomization tests, and full replication of data collection and analysis, with or without meaningful variations.

A. Comparison across split-halves of the data.
   1. When to split. Analysis of split-halves of one's data can be risky if one has only a few subjects to begin with. The half-size sample may be too small to reveal most of the interesting patterns. The number of subjects needed in each split-half depends on the reliability of the data and the number of dimensions one intends to try to verify. With most social sciences data (including rating scale data), a minimum of 10–15 in each half would be needed to verify a few of the largest dimensions. To more sensitively test for smaller dimensions, 35–70 in each split-half is preferable. With 100+ in each half, the method becomes a very powerful way of verifying consistencies of subtler relationships within each dimension or of estimating the reliability of less robust characteristics of the solution, such as the angle between dimensions. And with these larger split-halves, one can demonstrate the reliability of dimensions that do not contribute large portions of variance but may be theoretically interesting; it also allows one to reliably extract larger numbers of dimensions.
   2. How to split.
      a) In large data sets, random division into two groups should be sufficient. But with smaller data sets, one must guard against "unlucky" splits, where the two halves are actually (by chance) different. To guard against this, one can use the "orthogonal split-halves" technique. Divide the data randomly into four subsets: *A*, *B*, *C*, and *D*. Then construct the following alternative split-half divisions: (*A* + *B*) versus (*C* + *D*); (*A* + *C*) versus (*B* + *D*); and, if desired, (*A* + *D*) versus (*B* + *C*). Check each dimension or aspect of the solution for replicability across these two or three different splits. If the dimension replicates across any one of them, it is tentatively verified, since *any* such replication is very unlikely to happen by chance.
      b) If one is testing generalizability to new subject

samples, (perhaps the most common test), one splits the data into two subsamples of subjects. However, one can treat other modes of the data as representative samples from which one wishes to generalize. For example, one might sometimes split across variables, occasions, and so on. The question of which mode(s) to split is related to the "random effects versus fixed effects" question in analysis of variance. Those modes that you consider "random effects" —that is, those you consider simply a representative sample of possible levels and for which you would like to generalize any results to other similar samples of levels—should in theory be tested for generalizability by some technique such as split-half or bootstrapping.

3. How to measure agreement across split-halves. As noted at the beginning of this section, the most important objective of reliability measurement is usually to determine the generalizability of conclusions or interpretations; but this often calls for a difficult-to-quantify comparison of the conclusions that would be drawn from two different split-half solutions. More quantitative evaluation of the similarity of characteristics of solutions, such as patterns of loadings on a given dimension, can be computed by means of correlation coefficients or factor congruence measures. When loadings have a mean near zero (for instance, if they come from a mode that was centered), then correlations and cross-product congruence measures give the same result, but when the loadings are mostly positive (or negative), then the different measures of similarity stress different things. Cross-product measures of factor congruence are often preferred because they are sensitive to differences in overall elevation as well as profile shape of the loading pattern. However, cross-product measures of similarity can give very high and possibly misleading results when two factors with all positive loadings are being compared. One's choice should depend on which aspects of similarity are important. I usually recommend correlation as a more stringent test of factor similarity; it stresses those variations in loading size that are crucial to interpretation. However, there are circumstances in which correlations can also be misleading, so careful consideration is advised. (Insert C-1 goes here)

B. Resampling techniques for estimation of reliability: bootstrapping and jackknifing. Tukey's "jackknifing" and Efron's "bootstrapping" techniques are beginning to be used to measure the reliability of factor loadings and fit values. In these procedures, the original data set is used to generate several alternative versions, which are all analyzed and the results compared. The more

these alternative analyses differ, the less reliable is the conclusion based on the original data set.

1. Jackknifing. In this technique, new data sets are generated by omitting parts of the original data set. For example, a data set consisting of ratings made by 30 subjects could be used to construct 30 new data sets, each with one subject's data missing. Each of these data sets is then analyzed in the dimensionality under test, resulting in 30 sets of factor loadings. Reliability estimates are then computed from the variations across these solutions.
2. Bootstrapping. In this technique, new data sets are generated by sampling the original data *with replacement*. For example, if there are 30 subjects and generalizability across subjects is being tested, then new data sets of size 30 could be constructed, each of which is based on sampling of subjects *with replacement* from the original 30. Thus, in each of the new samples, some subjects will probably be omitted and others will occur more than once. Once again, the variability of loadings or fit values (or other parameters, such as angles between dimensions) across these new samples is used to estimate the reliability of loadings or fit values in the original data. (References for both bootstrapping and jackknifing are given in chapter 5.)

C. Cross-validation by applying loadings to a new sample. In some programs, such as PARAFAC and INDSCAL, it is possible to input Mode *A* and *B* loadings that have been determined by analysis of one sample and fit these to a new sample, estimating only new subject weights. Alternatively, PARAFAC allows one mode to be input and fixed and two to be estimated from the new sample. The resulting fit value will show "shrinkage" because you are no longer fitting error in two of the three modes. A sample can be split in half and the dimensions fit in each half can be applied to the subjects in the other half for a double cross-validation.

D. Randomization tests. "Significance tests" for the presence of system variation, additional dimensions, or other characteristics of the data can be obtained by means of randomization tests, in which the results of the analysis of the obtained data are compared with the results of randomly permuted versions of that data. For example, to test for system variation in a three-way array, one can randomly permute the entries in each "tube" of the data, thereby leaving the structure in two modes intact but scrambling the structure in the third mode. If the data are so permuted and analyzed 19 times, then under the null hypothesis, these 19 should not differ systematically from the original unpermuted data. Thus, the null hypothesis of no system variation can be rejected at the .05 level if the original unpermuted data produced a higher fit value than the 19 permuted ones, since this fit ranking would have a

probability of .05 under the null hypothesis. In a similar fashion, tests for additional dimensions can be constructed by permuting residuals from the lower dimensional analysis, tests for systematic asymmetries can be performed by randomly interchanging the x_{ij} and x entries in each allegedly symmetric data matrix, and so on.

E. Comparison across experiments. The most complete test of the generalizability of a given finding is, of course, when someone replicates the finding with a new sample, perhaps incorporating some modest variations in the methodology of data collection or subject selection, followed by a new analysis. Results that are stable across such replications demonstrate the strongest evidence for generalizability.

SUMMARY

Diagnostic evaluation of the optimality, reliability, and validity of solutions is often lacking in studies using multivariate methodology. Yet, the use of diagnostics is crucial because it enables the analyst to address four basic questions underlying any multivariate analysis: (a) appropriateness of the model; (b) computational adequacy of the fitting procedure; (c) statistical reliability of the solution; and (d) the generalizability and explanatory validity of any resulting interpretations.

A number of important diagnostic techniques for factor analysis and MDS are now available, including many that have been recently developed, and they could play an important role in promoting the growth and intelligent use of factor-analytic and MDS procedures over the next few years. To increase awareness of these techniques, an informal listing is presented of diagnostics known to the author (including some still being developed). In this listing, an attempt is made to develop a natural classification scheme for the diagnostics, based on the type of information they need (such as the data itself, the factor loadings resulting from a single analysis, loadings from several analyses, and so on) and which aspects of the solution they focus on (including the loadings, the residuals, and the fit values). Although detailed questions of how to use these diagnostics are not covered, brief descriptions of usage are sometimes given in the listing. Of the various techniques listed, the most important are probably the methods of evaluating the reliability of any characteristics of a solution (split-half, bootstrapping/jackknifing, and so forth). These can be used to estimate maximum dimensionality and decide which aspects of a solution are stable enough to warrant interpretation.