Cognitively Diagnostic Assessment

Edited by
Paul D. Nichols
University of Wisconsin, Milwaukee

Susan F. Chipman
U.S. Office of Naval Research

Robert L. Brennan
American College Testing

LEA LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
1995 Hillsdale, New Jersey Hove, UK

Chapter 13

Using Clustering Methods to Explore the Structure of Diagnostic Tests

James E. Corter
Teachers College, Columbia University

INTRODUCTION

The Nature of Diagnostic Tests

There are many reasons for the increase of interest in diagnostic testing. One reason is that well-designed diagnostic tests can provide very detailed assessments of achievement in academic domains. In such applications, the goal is not to provide a single number summarizing a student's geometry achievement, for example, but to separately assess the student's ability to answer various specific types of geometry problems, say, problems that deal with the congruence of two triangles or that involve applying a specific theorem. The results of such diagnostic tests might be used to design effective interventions for individual students, or be combined across students in order to evaluate particular curricula. In either case, the value of the diagnostic test arises from the fact that such a test provides not a single numeric index of overall skill of a student, but rather describes the student's specific pattern of mastered and nonmastered subskills.

Diagnostic tests may also prove useful to help validate specific models of human cognition. In recent years, cognitive scientists have provided fascinating descriptions of human problem solving in a variety of domains such as chess (Chase & Simon, 1973), multicolumn subtraction (Brown & Burton, 1978; VanLehn, 1982), computer programming (Adelson & Soloway, 1988; Anderson, Farrell, & Saurers, 1984; see also chapter 2 of this volume), and understanding mechanics problems in physics (Chi, Bassok, Lewis, Reimann, & Glaser, 1989;

VanLehn & Jones, 1993). Performance in such tasks is usually analyzed as being composed of very specific procedural subskills or subprocesses. It is challenging to try to develop measurement models that can be used to assess how well a student has mastered particular subskills. The use of such measurement models might also shed light on whether the proposed subskills have any psychological reality.

In order to understand what measurement models for diagnostic tests ought to look like, it is useful to consider some ways in which diagnostic tests differ from traditional tests of achievement or mental abilities. Diagnostic tests are designed to assess the state of mastery of very specific cognitive skills or operations. These subskills often have a discrete, "all-or-none" quality to them. For example, an arithmetic problem either does or does not require knowing how to subtract a negative number. And when a student attempts to solve a problem, she or he either does or does not perform this operation correctly (although across problems, a student may sometimes perform the subprocedure correctly and sometimes not). In contrast to the all-or-none nature of these very specific subskills, traditional tests are designed to measure more broadly defined mental abilities, and are based on the assumption that these abilities vary continuously along quantitative continua or "dimensions." That the diagnostic assessment task is a more difficult one is suggested by the possibility that two very different patterns of subskill mastery might both correspond to a single point on the traditional ability continuum. However, some test theorists would argue that such a finding indicates that the test domain does not correspond to a unitary skill, and that therefore a unidimensional model is not appropriate. A researcher might attempt to address this problem by proposing a multidimensional latent trait model for the domain.

Several other characteristics distinguish the diagnostic testing problem from more traditional testing situations. First, for many domains in which we need diagnostic tests (e.g., mathematics problem solving), a single problem might be solved by any of several valid solution paths. Thus, it may occasionally be difficult to determine from test data alone whether a student has acquired a particular subskill, or has simply chosen a different method of solution that does not require that subskill. Second, there may exist certain temporal or logical dependencies among subskills, such that subskill B cannot be acquired until after subskill A, or that subprocedure Y cannot be applied in the solution of a problem until subprocedure X has been successfully completed. These sorts of dependencies may make it more difficult to gather data that can be used to assess independently each of the subskills assumed to play a part in the problem-solving task. Table 13.1 summarizes some of the differences between diagnostic tests and traditional tests of mental abilities.

The idea that the cognitive attributes or subskills that we wish to measure often have a discrete, all-or-none character is central to the present chapter. The

TABLE 13.1
Some Relevant Characteristics of Diagnostic Tests and Traditional Tests of Mental Abilities

Diagnostic Tests	*Traditional Tests*
-measure very specific attributes	-measure broad abilities
-attributes may be "all-or-none"	-level of skill assumed to vary continuously
-subskills may be combined in any pattern	-unidimensional or multidimensional skill?
-different strategies may be used by different subjects or on different occasions	
-temporal or logical dependencies among subskills may exist	
-discrete-features models of item similarity seem appropriate (e.g., trees, clustering)	-geometric models of item similarity have been used (e.g., FA, PCA, MDS)

development of traditional mental tests was aided greatly by the use of dimensional models such as factor analysis (FA) and principal components analysis (PCA) to represent the correlations among items and subtests. These exploratory data-analysis methods have been used in many ways, for example, to explore the structure of tests being developed and to check the validity and reliability of a test using independent groups of subjects or with parallel forms of a test. The purpose of this chapter is to argue that due to the nondimensional, discrete character of the sort of cognitive attributes measured by diagnostic tests, alternative methods of exploring structure in proximity data (including correlations) are more appropriate for analyzing relations between the items of such tests than are the traditional techniques of FA and PCA (cf. Beller, 1990). Specifically, this chapter discusses the appropriateness of certain discrete-feature models of proximity for the representation of the correlations among diagnostic test items.

New Measurement Models for Diagnostic Tests

Several new measurement models specifically designed for diagnostic tests have recently been proposed. Fischer (1973) described a linear logistic test model (LLTM) for cognitive test items. In this approach, items are described both by item difficulty parameters similar to those used in item-response theory (IRT) models, and by a cognitive attributes matrix meant to represent the cognitive operations used in solving a particular item. This matrix (often termed the "Q" matrix) is a $K \times n$ matrix of 1s and 0s, where K is the number of presumed cognitive operations or "attributes" and n is the number of items. The (k,j)-th entry of this matrix is 1 if Item k requires the j-th attribute and 0 otherwise. In the LLTM, each cognitive attribute is also associated with a difficulty parameter. Embretsen (1984) proposed a general multicomponent latent-trait model combining the LLTM with a multicomponent latent-trait model.

In a study of mixed-fraction subtraction, Tatsuoka (1990; see also chapter 14 of this volume) used a proposed cognitive attributes or "Q" matrix to generate predicted "bugs" resulting from nonmastery of the individual cognitive attributes. Each of the resulting hypothetical bugs can be mapped, using the Q matrix, into a $1 \times n$ binary vector representing which items a student with the bug would get correct and which items the student would miss, assuming that the student made no other mistakes. This pattern of item performance is termed the "ideal" or "prototypical" pattern for that bug. Students were then classified, using the rule space methodology (Tatsuoka, 1985), according to which bug pattern their responses most closely resembled.

In chapter 15 of this volume, DiBello, Stout, and Roussos present what they term the "unified model." This model is designed to assess both a student's mastery of each of a set of cognitive attributes and a residual ability parameter interpreted as the student's ability to solve items outside of the cognitive framework represented by the Q matrix. The model, as mathematically formulated, can also represent the possibility that more than one strategy may be available to the student. Such an alternative solution strategy would be represented by an additional Q matrix.

Both the validity and the practicality of the new measurement techniques just described depend critically on the assumed set of cognitive attributes summarized in the Q matrix. Obviously, if the proposed cognitive attributes do not correspond to any real aspect of subjects' problem-solving behavior, then assessments of a subject based on these attributes will be meaningless. Furthermore, the omission of important attributes or the inclusion of superfluous ones may bias the estimation of parameters for the components that are present, just as can happen in multiple regression with a poorly chosen set of predictors. The inclusion of "too many" attributes can also lead to practical problems, though these may be relatively superficial in nature. For example, Tatsuoka (1990) initially investigated a Q matrix for the SAT Mathematics test consisting of 27 attributes for 60 items. This led to more than 3,000 prototypical bug patterns, which proved to be too many for the software used to analyze the patterns to be applied. Subsequent analyses, using a more parsimonious set of 14 attributes for 25 items only, produced a manageable set of 600 prototypical bug patterns.

Thus it is apparent that methods for assessing the validity and importance of proposed cognitive attributes for a set of items ought to be developed, in order for these new measurement models to provide maximally meaningful and useful results. The tree-fitting and nonhierarchical clustering methods described in this chapter may prove to be useful for this purpose. The techniques may be useful either as exploratory data analysis tools for investigating the structure of a diagnostic test for which no cognitive task analysis yet exists, or as confirmatory tools that could serve as an additional means of validating a proposed characterization of a test in terms of a set of specific cognitive attributes.

DISCRETE-FEATURE MODELS OF THE PROXIMITY AMONG TEST ITEMS

This section briefly introduces some of the basic types of discrete-feature models of similarity or proximity data that have been used in psychological research. Correlations among items can be considered as a form of proximity data, because the correlation between two items will presumably be higher if the two items test related knowledge or require the same cognitive operations; thus a higher correlation can be expected to correspond to a relatively more similar pair of items. The focus here is on hierarchical (tree) and nonhierarchical cluster models, rather than on spatial "geometric" models (FA, PCA, and multidimensional scaling) or on network models (see, e.g., chapter 10 of this volume) fit by such algorithms as NETSCAL (Hutchinson, 1989) or Pathfinder (Schvaneveldt, 1990) (see also Klauer & Carroll, 1989). Trees and nonhierarchical cluster models seem to offer certain advantages over the spatial and network models in representing the correlations among diagnostic test items. Chief among these advantages is the possibility of identifying parameters of the model (i.e., clusters and their weights) directly with the hypothesized cognitive attributes of a diagnostic test. As Tversky (1977) noted, a cluster comprised of k items may be interpreted as representing some attribute or feature shared by those k items. For diagnostic test items, such a shared attribute may be interpreted as a cognitive subskill or operation required by all of the items. Note that Beller (1990) compared the use of trees and spatial models to represent the relationships among test items, although not specifically in application to cognitively diagnostic tests. She found that for tests measuring achievement in multiple subject areas, trees did a better job of fitting the interitem correlation matrices than did spatial (MDS) models. Furthermore, distinct branches of the tree solutions corresponded to actual subtests (i.e., distinct subject areas) of the tests.

Hierarchical (Tree) Models of Proximity

Two basic types of tree models have been used in the social sciences to model proximity relations among conceptual "objects": the ultrametric tree (Johnson, 1967) and the additive tree (Sattath & Tversky, 1977). The ultrametric tree is the familiar "dendrogram" provided by the hierarchical clustering routines of many statistical packages. An example of an ultrametric tree representing the proximities (correlations) among five hypothetical test items (i1–i5) is shown in Fig. 13.1a. Note that there is a special distinguished point (located at the leftmost point in the diagram) in the tree graph known as the "root" of the tree. Note further that all "leaf nodes" in the tree (the points representing the items being clustered) are equidistant from this root; this is one of the distinguishing properties

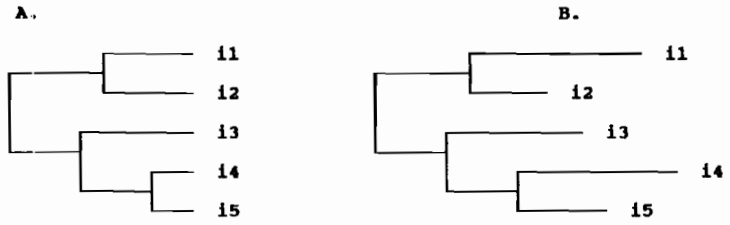

FIG. 13.1. A. An ultrametric tree representing the correlations among five items. B. An additive tree representing the correlations among five items.

of the ultrametric tree. Ultrametric trees are fit by the hierarchical clustering routines in many statistical packages; the "weighted average" version is recommended over the "single-link" or "complete-link" methods, for reasons of robustness (Corter, in press).

The restriction that all leaf nodes be equidistant from the root does not apply to additive trees (Fig. 13.1b). Thus an ultrametric tree is a special case of an additive tree in which this restriction holds. One implication of this fact is that not every proximity matrix that can be represented by an additive tree can be represented by an ultrametric tree. Also, the possibility of allowing items to differ in their distances from the root gives the additive tree the capability to represent the *uniquenesses* of items; that is, the extent to which items have unique characteristics not shared by any other item in the test. An item with relatively few unique characteristics will be relatively near to the root of the tree; a relatively unique item will be relatively far from the root. For example, in the additive tree of Fig. 13.1b, Item i2 is the least unique item, and Item i4 the most unique. Note that this concept of item uniqueness has a corresponding concept in common factor analysis. In factor analysis, item-specific variance is represented by the variable's uniqueness parameter (which is defined as one minus the variable's communality). This special capability of the additive tree seems particularly useful in representing the relations among test items. For a more thorough discussion of differences between ultrametric trees and additive trees, the interested reader is referred to Sattath and Tversky (1977) or Corter (in press). The most widely used program for fitting additive trees, ADDTREE/P, incorporates a modified version (Corter, 1982) of the algorithm described by Sattath and Tversky.

One of the attractive characteristics of trees as models of proximity data is the possibility of interpreting aspects of the tree solution in terms of the features or attributes shared by subsets of items. As is explained later, an ultrametric tree has a natural interpretation in terms of the common features of sets of items, whereas an additive tree can be understood in terms of the distinctive features of two sets of items (i.e., the features shared by the items in Set A that the items in Set B do not have, and conversely).

In a rooted tree, each arc in the tree determines a "branch" of the tree that corresponds to a particular set of items. The length of the arc can be interpreted

as indicating the importance of the features that are shared by that set of items (i.e., their common features). For example, in Fig. 13.1a there are three arcs that may be interpreted in this way: the arc above the node joining Objects i1 and i2, the arc above the node joining Objects i4 and i5, and the arc above the node that joins Object i3 to the node comprised of i4 and i5. Thus a rooted ultrametric tree can be thought of as representing the proximities among objects solely in terms of the *common features* shared by various sets of items (Sattath & Tversky, 1977). A tree model is not a completely general version of the common features model (Tversky, 1977), however, because only certain subsets of items (those subsets corresponding to branches or clusters in the tree structure) can share common features in the tree representation.

An additive tree too may be interpreted in terms of items and their features. In particular, an additive tree may be interpreted as representing the proximity between two items, x and y, in terms of their *distinctive features* (Tversky, 1977), that is, the features that x has that y does not, and the features that y has that x does not. Specifically, an arc in a rooted additive tree determines a branch of the tree corresponding to a specific set of items, and the length of the arc represents the total weight of the distinctive features of items in that cluster versus all other items in the total set. Thus there is some ambiguity in interpreting parameters in an additive tree—a given arc (connecting Subsets X and X^C of items, where X^C is the complement of X) may represent the weight of either features shared by Set X (but not by X^C), or features shared by Set X^C (but not by X), or some combination of the two.

In an additive tree, the placing of the root is arbitrary. A change in the rooting may affect the interpretation of the tree structure in terms of common features, because the change in rooting may change the branch determined by a specific arc from Set X to Set X^C. This would change the apparent interpretation of the arc from common features of Set A to common features of Set X^C. In fact, both interpretations are possible (because the placement of the root is arbitrary), thus one should be willing to consider either interpretation of the arc and its length.

To summarize, rooted trees may be interpreted as representing the common features of sets of items. When applied to a matrix of correlations among diagnostic test items, these common features may be identified with the common cognitive operations or subskills required to correctly answer the items. For example, the existence of a cluster consisting of Items 1, 3, and 7 of a test would provide evidence that these three items share one or more common cognitive components. The importance of this component in determining performance is indexed by the length of the arc corresponding to this cluster of items. Finally, it is argued that additive trees are more appropriate than ultrametric trees for analyzing test item correlations, because of the capability of the additive tree to represent item uniquenesses. However, the interpretation of a rooted additive tree is less straightforward than that of an ultrametric tree, because a given arc may represent either features of the set of items in that branch, or features shared by all other items not in the branch. An example of this nonuniqueness arises in one of the applications described later.

Nonhierarchical Clustering Models

Nonhierarchical clustering models also represent the similarity between two items in terms of the number and salience of the clusters to which they both belong. However, these models do not restrict the set of clusters to be nested (i.e., to form a tree or hierarchy). Two nonhierarchical cluster models have been widely used in psychology to represent the proximities among items: the ADCLUS model (Arabie & Carroll, 1980; Shepard & Arabie, 1979) and the extended tree or EXTREE model (Corter & Tversky, 1986).

The ADCLUS or additive clustering model (Shepard & Arabie, 1979) represents the similarity between two items x and y as the sum of the weights of each cluster of which x and y are both members. If we assume that each cluster represents a feature (cognitive component), then it is clear that the ADCLUS model is a common-features model. That is, the similarity between x and y is a simple sum of the weights of each feature (component) that x and y share. The program most commonly used to fit the ADCLUS model is the MAPCLUS program (Arabie & Carroll, 1980). This program provides a solution that consists of a set of clusters, each cluster specified as a set of items. Each cluster is associated with a weight indicating the cluster's importance. The user must specify how many clusters the program should find. Note that an ADCLUS solution provided by the MAPCLUS program closely resembles a cognitive attributes or "Q" matrix. Each cluster consists of a set of items that incorporate that attribute, and may be represented by a binary n-vector, with values of 1 indicating that that item incorporates the attribute and values of 0 indicating that it does not.

However, the ADCLUS model is a common-features model, meaning that it has no way of directly representing item uniquenesses. A nonhierarchical distinctive-features model that does have this capability is the extended tree or EXTREE model (Corter & Tversky, 1986). The graphical representation of the EXTREE model extends the notion of an additive tree by adding "marked segments" to represent clusters that "cut across" the basic tree structure (i.e., that do not form a nested set with the other clusters). An example of an extended tree is shown in Fig. 13.2. For example, the marked feature "A" occurs on two arcs, one leading to Item i3 and one leading to Items i7 and i8. This is meant to represent the fact that these three items share some feature (cognitive attribute). The distance between two items in an extended tree is calculated as the path length between them, except that if two items share a common marked segment (e.g., Items i3 and i7 share marked feature "A") then the lengths of the common marked segments do not enter into the path-length distance between those two items.

The extended tree preserves many of the graphical advantages of the additive tree representation. For example, the importance of clusters is indicated graphically by the lengths of arcs in the tree. Also, differential item uniquenesses due to item-specific content can be modeled by the lengths of the leaf arcs. But the extended tree offers the important additional benefit of being able to represent nonnested sets of cognitive attributes. Table 13.2 illustrates why this capability

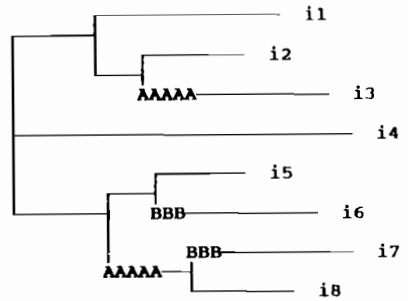

FIG. 13.2. An extended tree representing the proximities among eight items.

TABLE 13.2
Two Proposed Sets of Cognitive Attributes Corresponding to Two Different Methods of Solving Mixed-Fraction Subtraction Problems

Method A
(Always convert mixed numbers to improper fractions)

	Items																			
	01	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*09*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
A1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0
A2	0	0	0	1	1	0	1	0	1	1	1	0	1	1	0	1	1	1	1	1
A3	0	0	0	1	1	0	0	0	0	1	1	0	1	1	0	1	0	1	0	1
A4	1	0	0	1	1	0	0	0	0	1	1	1	0	0	0	0	0	1	1	1
A5	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
A6	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	1	1
A7	0	0	0	0	1	0	0	0	1	1	1	1	0	0	1	1	1	1	1	1

Method B
(Separate mixed numbers into whole and fraction parts)

	Items																			
	01	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*09*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*	*20*
B1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0
B2	0	0	0	1	1	0	1	0	1	1	1	0	1	1	1	1	1	1	1	1
B3	0	0	0	1	1	0	1	0	0	1	1	1	0	0	1	0	0	1	1	1
B4	1	1	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
B5	0	0	0	1	0	0	0	0	0	1	1	0	1	0	0	0	1	1	1	1
B6	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0
B7	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0

Note. From Tatsuoka (1990). Copyright 1990 by Lawrence Erlbaum Associates. Adapted by permission.

may be needed. This table presents two alternative sets of cognitive attributes (corresponding to two alternative solution strategies) proposed by Tatsuoka (1990) for a set of mixed-fraction subtraction items. For example, Attribute B1 corresponds to the subskill "Convert a whole number to a fraction or a mixed number" (Tatsuoka, 1990). The attributes in either set do not form a nested set or hierarchy, thus they cannot be represented by a tree structure (cf. Carroll & Corter, in press).

The relations between the four proximity models discussed here (ultrametric trees, additive trees, additive clustering, and extended trees) are summarized in Table 13.3. Both types of trees can represent only hierarchical or nested attribute structures, whereas the ADCLUS and EXTREE models can represent nonnested structures. Ultrametric trees and the ADCLUS model are best interpreted as common-features models (Tversky, 1977), whereas additive and extended trees can be interpreted in two ways. The first interpretation is in terms of distinctive features: The length of an arc connecting two subtrees represents the importance of the distinctive features of the two subsets of objects determined by the subtrees. The second interpretation is in terms of a set of common features plus item uniquenesses.

Thus it seems that the extended tree model may be especially appropriate for analyzing the correlational structure of diagnostic tests, because: (a) It is a discrete-feature model that represents interitem correlations in terms of a set of discrete attributes, (b) it has the capability to represent nonhierarchical or nonnested sets of cognitive attributes, and (c) it can allow for differential item uniquenesses caused by item-specific content.

APPLICATIONS

Applying the extended tree or EXTREE model to a matrix of correlations among diagnostic test items will result in a set of clusters of items being identified. Presumably, each cluster (or at least each highly weighted one) corresponds to

TABLE 13.3
A Classification of Some Discrete-Feature Models of Similarity

		Common- or Distinctive-Feature Model?	
		Common	*Distinctive*
Relationship among Feature Sets	Nested (Hierarchical)	ultrametric trees	additive trees
	Nonnested (Nonhierarchical)	additive clustering	extended trees

some cognitive attribute or attributes shared by the items in that cluster. Of course, some clusters might correspond to other types of shared characteristics of the items, for example, common domain content or even serial position in the test.

Such an analysis may be attempted either in an exploratory or a confirmatory spirit. If no cognitive task analysis of the test items has yet been attempted, then the EXTREE analysis might be conducted in an exploratory mode, to see what groups of items emerge in the extended tree structure. Each of these clusters could be examined to determine what, if any, cognitive attributes the items seem to share. On the other hand, if a characterization of a diagnostic test has already been made in terms of a proposed set of cognitive attributes (i.e., a specific Q matrix), then the EXTREE solution may serve as an additional means of validating this hypothesized attributes structure. Both of the applications reported next are essentially confirmatory, in that a specific Q matrix has either been proposed or is easily derivable from what is known about performance in the domain.

Mixed-Fraction Subtraction

The Q matrix labeled as "Method B" in Table 13.2 presents a set of cognitive attributes or subskills proposed by Tatsuoka (1990) to comprise one method of solving mixed-fraction subtraction problems. Actual test items written to correspond to the columns of this Q matrix are presented in Table 13.4. There are 40

TABLE 13.4
40 Mixed-Fraction Subtraction Items Used by Tatsuoka

1. $\frac{5}{3} - \frac{3}{4} =$	21. $\frac{8}{5} - \frac{5}{6} =$
2. $\frac{2}{4} - \frac{3}{8} =$	22. $\frac{5}{3} - \frac{5}{6} =$
3. $\frac{5}{6} - \frac{1}{9} =$	23. $\frac{5}{6} - \frac{1}{15} =$
4. $3\frac{1}{2} - 2\frac{3}{2} =$	24. $4\frac{1}{3} - 3\frac{4}{3} =$
5. $4\frac{3}{5} - 3\frac{4}{10} =$	25. $3\frac{2}{3} - 2\frac{2}{6} =$
6. $\frac{6}{7} - \frac{4}{7} =$	26. $\frac{3}{4} - \frac{2}{4} =$
7. $3 - 2\frac{1}{5} =$	27. $4 - 3\frac{1}{6} =$
8. $\frac{2}{3} - \frac{2}{3} =$	28. $\frac{3}{4} - \frac{3}{4} =$
9. $3\frac{7}{8} - 2 =$	29. $4\frac{8}{9} - 2 =$
10. $4\frac{4}{12} - 2\frac{7}{12} =$	30. $5\frac{3}{15} - 3\frac{8}{15} =$
11. $4\frac{1}{3} - 2\frac{4}{3} =$	31. $5\frac{1}{4} - 3\frac{5}{4} =$
12. $\frac{11}{8} - \frac{1}{8} =$	32. $\frac{13}{9} - \frac{1}{9} =$
13. $3\frac{3}{8} - 2\frac{5}{6} =$	33. $4\frac{4}{9} - 3\frac{5}{6} =$
14. $3\frac{4}{5} - 3\frac{3}{5} =$	34. $4\frac{5}{7} - 4\frac{3}{7} =$
15. $2 - \frac{1}{3} =$	35. $2 - \frac{1}{4} =$
16. $4\frac{5}{7} - 1\frac{4}{7} =$	36. $5\frac{7}{9} - 1\frac{5}{9} =$
17. $7\frac{2}{5} - \frac{4}{5} =$	37. $8\frac{1}{3} - \frac{2}{3} =$
18. $4\frac{1}{10} - 2\frac{8}{10} =$	38. $5\frac{1}{10} - 3\frac{4}{10} =$
19. $4 - 1\frac{4}{3} =$	39. $5 - 2\frac{3}{4} =$
20. $4\frac{1}{3} - 1\frac{2}{3} =$	40. $5\frac{1}{5} - 2\frac{7}{5} =$

items because each item type (as defined by the Q matrix) has two parallel items. For example, Items 1 and 21 are parallel items, both meant to instantiate the item type defined by the first column of Table 13.2. This test was administered to approximately 600 junior high school students, and their performance on items recorded. Detailed analyses of student performance on the tests were reported by Tatsuoka (1985, 1990). Some new analyses of these data are presented later.

As a first step in determining whether evidence for the validity of the proposed cognitive attributes structures of Methods A and B (see Table 13.2) can be found in the correlations among items, data from the two parallel items of each type were combined into single "doubled" items. For example, a subject's scores on Items i03 and i23 were summed, resulting in a score for "doubled" Item i03/i23 that could range from 0 (if the subject got both items incorrect) to 2 (if the subject got both items correct). This was done to obtain a smaller matrix (n = 20 rather than 40) and to increase the reliability of the analyzed scores.

The Pearson correlations among these "doubled" items were analyzed by the EXTREE program (Corter & Tversky, 1986), specifying the correlations as similarities data. The EXTREE algorithm proceeds in several stages. The first stage seeks to find the best-fitting additive tree for the data. In the second stage, all possible marked features are considered and the least-squares estimate of their lengths are obtained. In the third stage, the marked features with the largest weights are selected to be included in the model, and all parameters (arc lengths) are estimated simultaneously. All parameters less than an arbitrary cutoff value (which can be controlled by the user) are dropped, and the model reestimated, until all parameter estimates are above the threshold value.

The additive tree fit by the EXTREE program accounted for 91.1% of the variance in the correlations. For the extended tree (shown in Fig. 13.3), seven marked features were incorporated, and the resulting model accounted for 95.8% of the variance. This increase in R^2 is approximately 5%; thus the marked features selected by the EXTREE program noticeably improve the fit of the model, compared with the additive tree. This confirms that, as expected, the test attributes determining the interitem correlations do not exhibit a hierarchical or tree structure.

Several interpretable clusters of items appear in Fig. 13.3. For example, at the top of the tree Items i01/21, i02/22, i03/23, and i05/25 are grouped into the same tree cluster. With the addition of Items i13/33, this cluster corresponds to cognitive attribute B4 (or A5). Although Item i13/33 is not included in this tree cluster, there is a marked feature (labeled "D") that joins i13/33 to a member of this cluster. Thus the marked features in some sense attempt to "repair" the tree structure, by finding and marking places where the patterns of similarity "cut across" the tree structure. Another interpretable cluster consists of Items i07/27, i15/35, and i19/39. These items are similar in that they are the only ones that require attributes B1, B2, *and* B3 (no interpretation of this cluster in terms of the Method A matrix was evident). Finally, at the bottom of the tree is a cluster

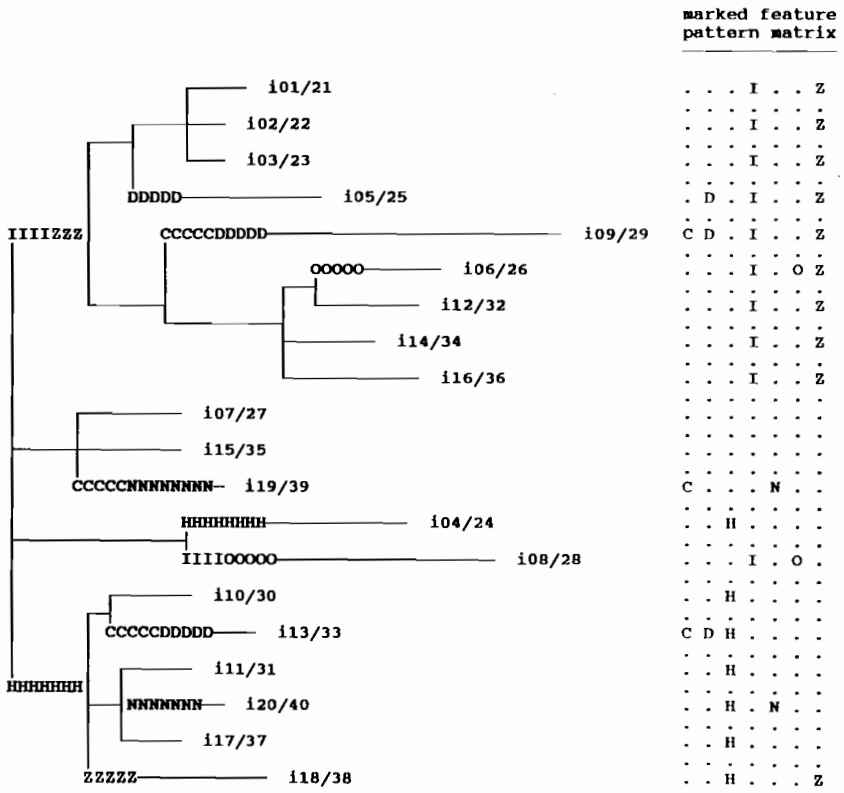

FIG. 13.3. Extended tree solution for the correlations among 20 "doubled" mixed-fraction subtraction items.

consisting of Items i10/30, i13/33, i11/31, i20/40, i17/37, and i18/38. This cluster, with the addition of Items i04/24 and i19/39, corresponds to attribute B5. One of these "missing" items, i04/24, is linked to this cluster by the marked feature "H." The other missing item, i19/39, is linked to two members of this cluster by the marked features "C" and "N." This cluster, with or without repairs, did not seem to correspond to any Method A attributes.

Thus the extended tree solution provides some support for the proposed set of Method B cognitive attributes for these items. However, the correspondence between clusters of the solution and specific proposed cognitive attributes is not perfect. For example, no evidence emerges from the tree solution supporting the existence of Attributes B6 or B7.

Additional analyses were performed to check the validity of these conclusions, namely that Attributes B1, B2, B3, B4, and B5 are relatively more predictive of item performance than are Attributes B6 and B7, and that the Method A attributes

seem to be relatively unrelated to item performance. These additional analyses used a method similar to one proposed by Scheiblechner (1972) for validating a proposed attributes matrix (described in Tatsuoka, 1990). In this method, one treats the rows of the Q matrix as predictor variables, with a value of 1 if the item in question involves that attribute and 0 otherwise. The proportion correct associated with each column (i.e., item) of the matrix is used as the dependent variable (Scheiblechner and Tatsuoka used item difficulties estimated using IRT models). Thus the proportion correct for each item is predicted as the sum of the weights (regression coefficients) of the variables (i.e., cognitive attributes) that enter into the item. The results of this analysis are shown in Table 13.5. The top half of the table reports the regression results for the Method A matrix (i.e., using those cognitive subskills presumed by Tatsuoka, 1990, to play a role in the proposed Method A for solving mixed-fraction subtraction problems), whereas the bottom half reports the Method B results (i.e., for the set of attributes proposed for the alternative solution strategy B).

The first thing to observe about Table 13.5 is that the Method B attributes matrix accounts for a much higher percentage of the variance in the item proportions correct than does the Method A matrix ($R^2 = .88$ for Method B, $R^2 = .52$ for Method A). The overall equation for the Method B matrix is significant ($F = 12.69$, $p < .001$), whereas for the Method A matrix it is not ($F = 1.846$, $p = .168$). Note that the meaningfulness of interpreting these and other p values from these regression analyses can be justified by a permutation-test interpretation of the F test (Freedman & Lane, 1983), even though the "observations" of this analysis (the items) cannot be assumed to be independent. The results indicate that the Method B matrix does a better job of predicting the difficulty of items, as measured by item percent correct. Another thing to notice in Table 13.5 is the sign of the coefficients of variables (attributes) in the multiple regression equation for Method B. The signs of all coefficients (except the one corresponding to B7) are negative, indicating that presence of the attribute predicts lower item percent correct. This is consistent with the expected direction of the relationship, which is based on the reasoning that the more cognitive attributes are necessary to solve the item, the more difficult it is. The one exception, the coefficient corresponding to B7, has the least significant p value attached to it ($p = .827$), therefore little confidence can be placed in any conclusion about its true sign. Finally, the two least significant Method B attributes in predicting item performance were B7 and B6. This too is consistent with the extended tree solution, which provided no evidence to support a role of these attributes in determining item performance.

Multicolumn Subtraction

VanLehn (1982) reported studies in which data were collected from several hundred elementary school students concerning their performance on multicolumn subtraction problems, using a number of different test forms. For the present

TABLE 13.5
Results of Multiple Regressions Predicting Proportions Correct for Items Using Two Proposed Attributes Matrices

Method A Attributes:
Dependent Variable: PCORR (Proportion Correct)
Squared Multiple R = .518

Variable	*Coefficient*	*Std Error*	*T*	*p(Two-Tail)*
Constant	.744	.073	10.195	.000
A1	−.154	.110	−1.402	.186
A2	−.169	.110	−1.539	.150
A3	.027	.112	0.237	.817
A4	−.098	.076	−1.296	.219
A5	−.193	.094	−2.066	.061
A6	.045	.086	0.528	.607
A7	−.025	.075	−0.337	.742

Analysis of Variance

Source	*Sum-of-Squares*	*DF*	*Mean-Square*	*F Ratio*	*p*
Regression	.216	7	.031	1.846	.168
Residual	.201	12	.017		

Method B Attributes:
Dependent Variable: PCORR (Proportion Correct)
Squared Multiple R = .881

Variable	*Coefficient*	*Std Error*	*T*	*p(Two-Tail)*
Constant	.750	.034	22.333	.000
B1	−.181	.049	−3.720	.003
B2	−.040	.044	−0.898	.387
B3	−.045	.044	−1.022	.327
B4	−.202	.038	−5.319	.000
B5	−.212	.045	−4.747	.000
B6	−.037	.044	−0.842	.416
B7	.012	.054	0.224	.827

Analysis of Variance

Source	*Sum-of-Squares*	*DF*	*Mean-Square*	*F Ratio*	*p*
Regression	.367	7	.052	12.689	.000
Residual	.050	12	.004		

investigation, only responses from Test Form 1 were compiled and analyzed (because this form had the largest number of respondents, $N = 520$). Form 1 consisted of 17 multicolumn subtraction items. These items varied in difficulty; the proportions of subjects getting each one correct varied from .276 (for Item 17, "10,013 – 318") to .794 (for Item 09, "654 – 204").

Much is known about the types of bugs that students exhibit in solving such problems (Brown & Burton, 1978; VanLehn, 1982, 1990). By carefully considering these bugs and the formal structure of the subtraction problems, it is possible to derive a set of procedural subskills (one specific type of cognitive attribute) hypothesized to play a part in solving such subtraction problems. Table 13.6 presents this hypothesized cognitive attributes or Q matrix. There are five proposed attributes or subskills, "subtract," "borrow," "double-borrow," "subtract-from-0," and "borrow-from-0." "Double-borrow" refers to the subskills needed to handle problems in which a borrow occurs in two adjacent columns (e.g., the 1s and 10s columns).

An attempt was made to validate this cognitive attributes structure by using the rows of the matrix in Table 13.6 as variables to predict item proportions correct, as was done with the previous example concerning mixed-fraction subtraction. However, because attribute AT1, "subtract," is required for all the items, its binary representation in the Q matrix is as a vector of all 1s; thus it is a constant and cannot be used as a predictor variable. The results of the regression analysis using attributes AT2–AT5 are presented in Table 13.7. The signs of the coefficients for all the variables are negative, indicating that presence of each attribute for the item leads to lower item performance (i.e., greater item difficulty), as expected. The most significant variable is AT2, "borrow," whereas the least significant is AT4, "subtract-from-0."

From the test data on the 520 subjects, Pearson correlations were computed among all pairs of test items (items left blank were coded as incorrect rather than missing). This correlation matrix was analyzed using the EXTREE program. In the first stage of the algorithm, an additive tree accounting for 97.6% of the variance was fit to the data. Four marked features were then fit by the algorithm, and the resulting extended tree accounted for 98.3% of the variance. The extended tree solution is shown in Fig. 13.4. An increase of less than 1% in variance accounted for with four marked features is not particularly impressive. Therefore it is concluded that the extended tree does not constitute any important improvement over the additive tree. Thus the additive tree solution for these data could be used for interpretational purposes, if desired.

At the bottom of the tree a particularly important cluster emerges, consisting of Items i05, i06, i12, i17, i14, and i15 (the importance of a cluster is indicated by the length of the arc "above" the cluster). These are exactly the items that require the subskill "borrow-from-0." Another salient cluster consists of only the two items i09 and i11. These are the only two items in the set that do *not* require the subskill "borrow." This result points up one of the potential pitfalls of inter-

TABLE 13.6
A Proposed Cognitive Attributes Matrix for Multicolumn Subtraction, Test Form 1 Items

	Item																
Attribute	*01*	*02*	*03*	*04*	*05*	*06*	*07*	*08*	*09*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*
AT1 SUBTRACT	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AT2 BORROW	1	1	1	1	1	1	1	1	0	1	0	1	1	1	1	1	1
AT3 DOUBLE-BORROW	0	0	0	1	1	1	1	1	0	1	0	1	0	1	1	0	1
AT4 SUBTR_FROM_0	0	1	0	0	0	1	1	1	0	0	0	0	0	1	0	1	1
AT5 BORROW_ FROM_0	0	0	0	0	1	1	0	0	0	0	0	1	0	1	1	0	1

TABLE 13.7
Results of Multiple Regression Predicting Item Proportions Correct From Five Cognitive Attributes Proposed for Multicolumn Subtraction

Dependent Variable: PROPCORR (Proportion Correct)
The following variables are constants or have missing correlations: AT1 Subtract
Squared Multiple R = .854

Variable	*Coefficient*	*Std Error*	*T*	*p(Two-Tail)*
Constant	.862000	.057247	15.058	.0000
AT2	−.236216	.069796	−3.384	.0054
AT3	−.116304	.054472	−2.135	.0541
AT4	−.062459	.042089	−1.484	.1636
AT5	−.128583	.052259	−2.461	.0300

Analysis of Variance

Source	*Sum-of-Squares*	*DF*	*Mean-Square*	*F Ratio*	*p*
Regression	.46146	4	.11536	17.60116	.0001
Residual	.07865	12	.00655		

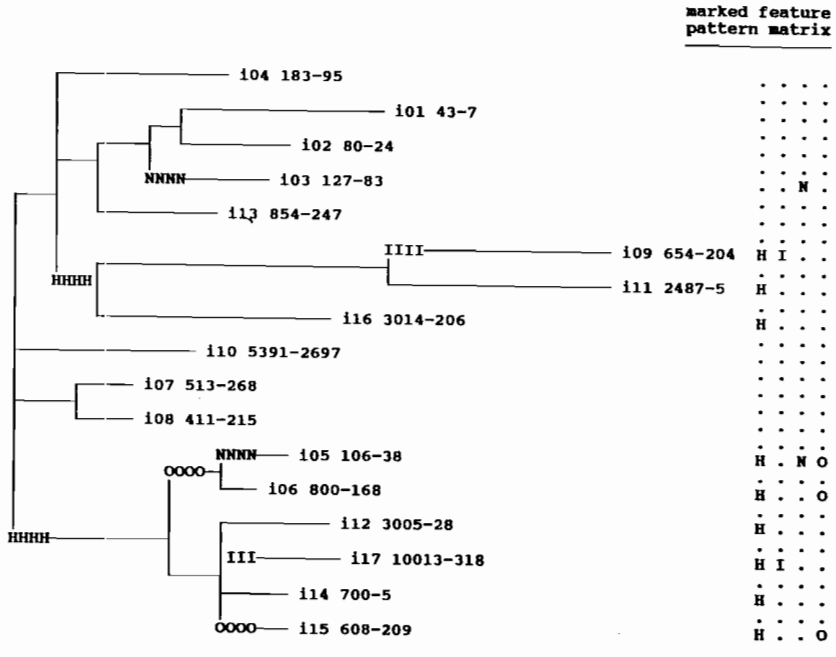

FIG. 13.4. Extended tree solution for the correlations among 17 multicolumn subtraction items.

preting the structure of an additive or extended tree solution: Because these are distinctive-features models (Sattath & Tversky, 1977), the user must be alert to the possibility that it is the *complement* of the set of items in a cluster, rather than the set of items actually in the cluster, that share some important attribute or subskill. Finally, less clear-cut evidence emerges from Fig. 13.4 to support the "double-borrow" attribute: The large cluster at the top of the tree consisting of Items i04, i01, i02, i03, i13, i09, i11, and i16 is roughly the set of items that does *not* require the skill of double-borrowing. The only exception is that Item i04 should not be associated with this cluster, under this interpretation. However, it is possible that Item i04 had a fairly high association with this cluster due to its serial position in the test immediately following Items i01–i03. Perhaps this anomalous membership would not have occurred if several different subforms of the test had been constructed with different orderings of items. No evidence for Attribute AT4, "subtract-from-0," is found in the tree solution. Again, this is consistent with the results of the regression analysis, which found this attribute to be least important in predicting item percent correct. It is interesting to note that omission of Attribute AT4 from the Q matrix of Table 13.6 results in a hierarchical structure: That is, all remaining attributes (AT1, AT2, AT3, and AT5) form a nested set in which each pair of clusters is either disjoint or one cluster is nested inside the other. This observation is consistent with the finding that the addition of (nonnested) marked features did not significantly improve the ability of the tree to account for the item correlations.

DISCUSSION

The discrete-feature models of proximity fit by certain clustering methods have been advocated in this chapter as being particularly well suited to exploring the structure of diagnostic tests, due to the presumed discrete nature of many cognitive "attributes" or subskills. These methods may be useful tools to develop, sharpen, and validate hypothesized sets of cognitive attributes. Such tools are needed, because the practicality and validity of many recently developed measurement models for diagnostic tests depend critically on the validity of the assumed attributes matrix.

Additive and extended trees in particular have been recommended here for representing the structure of item correlations for diagnostic tests. The chief advantage of additive trees over ultrametric trees, and extended trees over the additive clustering model, lies with the capability of additive and extended trees to represent item "specificities," or unique features. These item uniquenesses will often be interpretable in their own right, and their explicit representation in the model may lead to more interpretable common-feature structure (this is the logic underlying common-factor analysis).

In addition, it has been argued that the cognitive attribute structure of a set of diagnostic test items may have a nonhierarchical structure. Two such nonhier-

archical structures for sets of subtraction items were presented here. In such cases, a nonhierarchical clustering model, such as EXTREE (Corter & Tversky, 1986) or ADCLUS (Arabie & Carroll, 1980; Shepard & Arabie, 1979) will be needed to represent the attribute structure.

The general problem of trying to detect the use of a particular solution strategy when more than one strategy is possible deserves some additional comments. Note that the existence of two alternative solution strategies, as in the Tatsuoka (1985, 1990) data for mixed-fraction subtraction items, complicates the problem of cognitive diagnosis considerably. Although little evidence was found in the extended tree and regression analyses to support the importance of solution Method A, it may be that some (but relatively few) students were in fact using this strategy. The practical effect of these students' data being mixed in with data from relatively many "Method B" students may be to add "noise" to a data structure that would otherwise more strongly support the use of the attributes proposed for Method B. Even more worrisome is the possibility that an individual student may switch back and forth between the two solution strategies for different problems, which would also make it more difficult to detect the structures that EXTREE is being used to search for. Thus these data may not constitute the simplest or fairest test of the proposed method. A simpler test might be set up by an experiment requiring one group of subjects to solve the problems by Method B, and another group by Method A. This would result in two matrices of relatively "pure" correlation data, which could be used to test how well the extended tree method can detect the effects of particular attributes in a single-strategy application.

Given the difficulty of the present task, which is to detect the effects of particular attributes in data that may be noisy or generated by students using a mix of solution strategies, some closing comments may be in order concerning the inferences that can be made concerning the cognitive attributes structure of a set of diagnostic test items when positive or negative results are obtained for particular attributes. If a specific attribute has been hypothesized to play a part in solving certain items in a test, and that set of items emerges as a cluster in the extended tree solution, then the validity and importance of that attribute is supported. If, on the other hand, no cluster emerges corresponding to the set of items requiring a hypothesized attribute, then at least three possible explanations exist. One possibility is that the hypothesized attribute is not valid, or is not actually involved in solution of the items. A second possibility, already discussed, is that the attribute plays a part in one but not all alternative solution strategies that can be used for the items. A third possibility is that the attribute *does* play a role, as hypothesized, in the solution of the items, but does not predict a significant proportion of the variance in test performance. This could happen if the attribute in question has already been mastered by all of the subjects in the test population, for example. In this case, the attribute would not lead to higher correlations between items involving that attribute, and the cluster would not be found by any type of analysis of the correlation matrix. Attribute AT4, proposed for the VanLehn (1982) subtraction data, provides an example of this interpretational ambiguity. Although it certainly seems true that in order to solve Items i01, i06, i07, i08, i14, i16, and i17, one must be able to subtract from 0 (AT4), no evidence emerged from the extended tree analysis supporting the existence of this cognitive attribute. However, we cannot be certain if this means that there is no real cognitive subskill that corresponds to the particular operation of subtracting from 0 (as distinct from other subskills such as borrowing), or if there is such a particularized subskill but it has been mastered by such a high proportion of the population that the attribute has little value in predicting the correlations among test items.

Thus, null findings resulting from use of the present method to validate a prior cognitive model for a test domain seem susceptible to multiple interpretations. Further data, from alternate test forms or different subject populations, may be required to settle upon a single interpretation of a null result for a particular attribute. Depending on the strength of our prior belief in the model and on our pragmatic goals in the diagnostic testing situation, we might react to a null result by either redoubling our efforts to develop effective items to measure the attribute (if we are very sure that it exists), or by simply concluding that the attribute is not important enough to try to diagnose, at least for the populations of subjects and items being studied.

ACKNOWLEDGMENTS

I wish to thank Kikumi Tatsuoka and Kurt VanLehn for their generosity in making available their datasets on student performance in subtraction problems. Thanks also to H. Jane Rogers for serving as a resource regarding current research in item response theory, and to Susan Chipman, Paul Nichols, and Robert Brennan for the invitation to participate in the ONR/ACT Conference on Alternative Diagnostic Assessment.

Information on ordering the ADDTREE/P and EXTREE programs described in this chapter is available from James E. Corter, Box 41, Teachers College, Columbia University 10027.

REFERENCES

Adelson, B., & Soloway, E. (1988). A model of software design. In M. Chi, R. Glaser, & M. Farr (Eds.), *The nature of expertise* (pp. 185–208). Hillsdale, NJ: Lawrence Erlbaum Associates.

Anderson, J. R., Farrell, R. G., & Saurers, R. (1984). Learning to program in LISP. *Cognitive Science, 8*, 87–129.

Arabie, P., & Carroll, J. D. (1980). MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika, 45*, 211–235.

Beller, M. (1990). Tree versus geometric representation of tests and items. *Applied Psychological Measurement, 14*, 13–28.

Brown, J. S., & Burton, R. B. (1978). Diagnostic models for procedural bugs in basic procedural skills. *Cognitive Science, 2*, 155–192.

Carroll, J. D., & Corter, J. E. (in press). A graph-theoretic method for organizing nonhierarchical clusters as trees or extended trees. *Journal of Classification.*

Chase, W. G., & Simon, H. A. (1973). Perception in chess. *Cognitive Psychology, 4*, 55–81.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13*, 145–182.

Corter, J. E. (1982). ADDTREE/P: A PASCAL program for fitting additive trees based on Sattath & Tversky's ADDTREE algorithm. *Behavior Research Methods and Instrumentation, 14*, 353–354.

Corter, J. E. (in press). *Tree models of similarity and association.* Beverly Hills, CA: Sage.

Corter, J. E., & Tversky, A. (1986). Extended similarity trees. *Psychometrika, 51*, 429–451.

Embretsen, S. E. (1984). A general latent trait model for response processes. *Psychometrika, 49*, 175–186.

Fischer, G. H. (1973). The linear logistic test model as an instrument in education research. *Acta Psychologica, 37*, 359–374.

Freedman, D., & Lane, D. (1983). A nonstochastic interpretation of reported significance levels. *Journal of Business & Economics Statistics, 1*, 292–298.

Hutchinson, J. W. (1989). NETSCAL: A network scaling algorithm for nonsymmetric proximity data. *Psychometrika, 54*, 25–51.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika, 32*, 241–254.

Klauer, K. C., & Carroll, J. D. (1989). A mathematical programming approach to fitting general graphs. *Journal of Classification, 6*, 247–270.

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

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben [Learning and solving complex mental problems]. *Zeitschrift für Experimentelle und Angewandte Psychologie, 19*, 476–506.

Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review, 86*, 87–123.

Tatsuoka, K. K. (1985). A probabilistic model for diagnosing misconceptions in the pattern classification approach. *Journal of Educational Statistics, 12*, 55–73.

Tatsuoka, K. K. (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, & M. G. Shafto (Eds.), *Diagnostic monitoring of skill and knowledge acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327–352.

VanLehn, K. (1982). Bugs are not enough: Empirical studies of bugs, impasses and repairs in procedural skills. *The Journal of Mathematical Behavior, 3*, 3–71.

VanLehn, K. (1990). *Mind bugs: The origins of procedural misconceptions.* Cambridge, MA: MIT Press.

VanLehn, K., & Jones, R. M. (1993). Learning by explaining examples to oneself: A computational model. In S. Chipman & A. L. Meyrowitz (Eds.), *Foundations of knowledge acquisition: Cognitive models of complex learning* (pp. 25–82). Boston, MA: Kluwer Academic.