Conditional Covariance-Based Nonparametric Multidimensionality Assessment

William Stout and Brian Habing, University of Illinois
Jeff Douglas, University of Wisconsin
Hae Rim Kim, Sanji University
Louis Roussos, Law School Admission Council
Jinming Zhang, Educational Testing Service

According to the weak local independence approach to defining dimensionality, the fundamental quantities for determining a test's dimensional structure are the covariances of item-pair responses conditioned on examinee trait level. This paper describes three dimensionality assessment procedures—HCA/CCPROX, DIMTEST, and DETECT—that use estimates of these conditional covariances. All three procedures are nonparametric; that is, they do not depend on the functional form of the item response functions. These procedures are applied to a dimensionality study of the LSAT, which illustrates the capacity of the approaches to assess the lack of unidimensionality, identify groups of items manifesting approximate simple structure, determine the number of dominant dimensions, and measure the amount of multidimensionality. *Index terms: approximate simple structure, conditional covariance, DETECT, dimensionality, DIMTEST, HCA/CCPROX, hierarchical cluster analysis, IRT, LSAT, local independence, multidimensionality, simple structure.*

A fundamentally important problem for psychological and educational measurement is the item level statistical assessment of multidimensionality using item response data of a group of examinees to a set of test items. Current emphases on computerized adaptive testing, performance assessment, and cognitive diagnosis only strengthen the need for accurate multidimensional assessment tools. Assessment of test dimensionality usually occurs in two distinct operations: (1) the verification or refutation of unidimensionality, and (2) if necessary, the subsequent description of the test's multidimensional structure.

The verification of unidimensionality is often necessary because many currently used psychometric procedures applied to test data presume that the data fit a unidimensional latent model. Hence, use of these procedures (such as BILOG) can only be adequately justified by a statistical analysis confirming approximate unidimensionality, or more generally by a statistical robustness argument asserting that the amount of departure from unidimensionality does not seriously invalidate the use of a specific procedure.

If the hypothesis of unidimensionality is rejected, detailed information about the multidimensional latent test structure is often of substantive or methodological interest, either for psychological research or educational measurement purposes. The most basic type of latent multidimensionality is "simple structure." A test exhibits approximate simple structure if its items can be partitioned into item clusters that are each relatively dimensionally homogeneous and that are dimensionally distinct from one another. In this case, the number of clusters is equal to the number of dominant dimensions. For example, if the specifications of a mathematics test call for algebra and geometry items, a dimensionally homogeneous algebra cluster and a dimensionally homogeneous geometry cluster that together produce a latent structure with two dominant dimensions may

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 20, No. 4, December 1996, pp. 331–354

0146-6216/96/040331-24$2.45

be expected. It is important to note that simple structure does not imply that the dimensions are uncorrelated.

The classical approach to item-level dimensionality analysis has been to use parametric linear factor analysis on an item × item observed covariance matrix, possibly based on tetrachoric correlations. Although there is general agreement that this item-level linear factor analysis can perform poorly (e.g., Hulin, Drasgow, & Parsons, 1983; McDonald, 1981), several promising parametric nonlinear factor analysis approaches have been proposed as alternatives. Two such approaches are the limited information, weak local independence (WLI) based NOHARM approach (DeChamplain, 1995a; Fraser, 1988; Knol & Berger, 1991; McDonald, 1982) and the full-information strong local independence (SLI) based TESTFACT marginal likelihood approach (see Bock, Gibbons, & Muraki, 1988).

As opposed to these parametric approaches, the dimensionality assessment procedures described here are nonparametric. That is, no particular parametric form for the item response functions is assumed. In addition to being nonparametric, the procedures—HCA/CCPROX, DIMTEST, and DETECT—are all based on estimating the conditional covariances of the item pairs, where the conditioning variable is an appropriately selected subscore.

Item Pair Conditional Covariance

Let $\mathbf{U}_n$ denote the item response pattern of a randomly sampled examinee on a test of length n, and $\boldsymbol{\Theta}$ be the possibly multidimensional latent variable that underlies the item response pattern's distribution. $\boldsymbol{\Theta}$ is used here to indicate both a multidimensional latent variable and the coordinate system corresponding to that variable in its geometric representation. $\boldsymbol{\theta}$ will represent a particular value of $\boldsymbol{\Theta}$. Θ will represent both a unidimensional latent variable and the axis corresponding to that variable when represented using the coordinate system $\boldsymbol{\Theta}$. θ will represent a particular value of Θ. For convenience, $\boldsymbol{\Theta} = \boldsymbol{\theta}$ will often be denoted by $\boldsymbol{\theta}$. Similarly, for a unidimensional latent variable Θ, $\Theta = \theta$ will often be denoted θ.

Following the work of Mokken (1971), Stout (1990) defined the dimensionality d of a test $\mathbf{U}_n$ as the minimal dimensionality required for $\boldsymbol{\Theta}$ to produce a model that is both locally independent and monotone. Here, monotonicity is the condition that $P[U_i = u_i | \boldsymbol{\Theta} = \boldsymbol{\theta}]$ is increasing coordinate-wise in $\boldsymbol{\theta}$ for each item i. SLI is the condition that

$$P(\mathbf{U}_n = \mathbf{u}_n | \theta) = \prod_{i=1}^{n} P(U_i = u_i | \theta) \tag{1}$$

for each response pattern $\mathbf{u}_n$ and all $\boldsymbol{\theta}$. WLI is the condition that, for all $[n(n-1)/2]$ item pairs and all $\boldsymbol{\theta}$,

$$\mathrm{cov}(U_i, U_l | \theta) = 0, \tag{2}$$

where cov is the covariance.

SLI implies WLI by

$$\mathrm{cov}(U_i, U_l | \theta) = P(U_i = u_i, U_l = u_l | \theta) - P(U_i = u_i | \theta) P(U_l = u_l | \theta), \tag{3}$$

and then noting that Equation 1 implies

$$P(U_i = u_i, U_l = u_l | \theta) = P(U_i = u_i | \theta) P(U_l = u_l | \theta), \tag{4}$$

Thus, WLI is also referred to as pairwise local independence.

Many psychometricians (see McDonald, 1994) argue that, in cases of real test data for which WLI holds, SLI holds approximately. Assuming this to be true, and assuming monotonicity, testing for unidimensionality is then equivalent to testing for WLI; that is, testing that the conditional covariance is 0 for all values of some appropriately selected unidimensional θ and all item pairs.

If a test, or subtest, is unidimensional, then the particular scale used for θ is arbitrary because all strictly monotone rescalings of it produce the same unidimensional ordinal item response theory model. In particular, expected number correct (ENC; sometimes called "true score") is a monotone transformation of the θ scale selected for the unidimensional model. Because the number-correct (NC) score is an easily calculated consis-

tent estimator of ENC, the ENC scale is used here as the metric for examinee trait levels, and NC is substituted for θ when computing the estimates of the item-pair conditional covariances. Note, however, that the results presented here are unaffected by converting the ENC scale to other scales, such as the standard normal or percentile scale, by application of the appropriate monotone transformation.

In cases in which multidimensionality is present, NC is still used as the empirical conditioning variable. In this case, NC can be informally considered to be a consistent estimator of Θ_{TT}, the unidimensional latent variable "best measured" by the total test NC score. Here Θ_{TT} should be viewed as a direction or axis embedded in the multidimensional coordinate system $\boldsymbol{\Theta}$. A similar notation is used for subtests, where Θ_C denotes the unidimensional latent variable best measured by NC on subtest C. By combining the WLI definition of dimensionality and the consistency of NC score as an estimator of ENC, the empirical conditional covariance $[\widehat{\mathrm{cov}}(U_i,U_l \mid S_{i,l}]$ should be near 0 when unidimensionality holds and should be nonzero when it fails. $S_{i,l}$ denotes the NC score on the test obtained by excluding items i and l. Thus, these empirical conditional covariances form the basis for dimensionality assessment. The consistency of NC as an estimator of trait level as measured by ENC is discussed in more detail in Stout (1990). In the multidimensional case, one method of rigorously defining the concept of "best measured" latent variable is to use Wang's reference composite (see Wang, 1988). Another method is described in Zhang & Stout (1996b).

Geometric Representation of Multidimensional Tests

The geometric view of multidimensional latent variable models developed by Reckase, and used by Ackerman in an article in this issue (Ackerman, 1996; see also Ackerman, 1994), will serve as a useful intuitive guide. An example of this representation of items as vectors is shown in Figure 1.

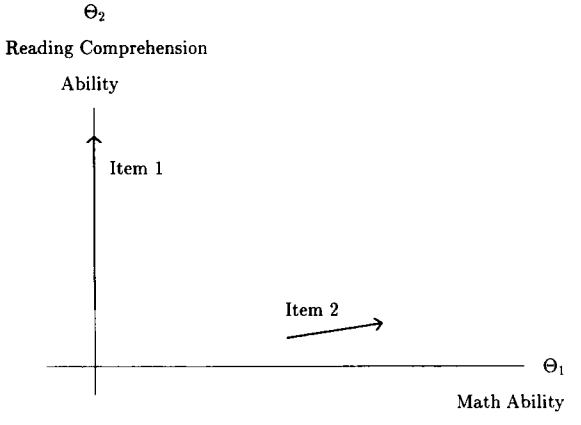
In Figure 1, the item response depends on the examinee's capacity in two traits: Θ_1 and Θ_2. Item 1 is a moderately challenging and highly discriminating reading comprehension item; Item 2 is a difficult mathematics story problem that requires only basic reading skills and is of moderate discrimination. The vector representing the item is selected from the class of vectors that lie on lines through the origin, where the origin is taken as the population multidimensional trait level mean. The direction of the vector is that in which the item has maximum discrimination (which must be well-defined in the context of the model being used). This direction is referred to as the item's direction of best measurement. The location of the base of the vector is such that an examinee of that multidimensional trait level would have .5 probability of correctly answering the item. The length of the vector is a measure of the magnitude of the item's discrimination. For the special case of a two-dimensional compensatory logistic model, the vector representing the item has the same direction and length as the item's discrimination vector.

The graphical examples used here will, for clarity, use items that are all of average difficulty and that have similar discriminations. The principles illustrated by the graphs also hold true for more graphically complex situations, including those occurring in higher dimensions.

Several concepts described above can be illustrated with the geometrical representation in Figure 1. First, (exact) simple structure is defined to exist for a d-dimensional test if a d-dimensional latent coordinate system $\boldsymbol{\Theta} = \{\Theta_i: 1 \le i \le d\}$ exists such that all items lie along the coordinate axes. Then, each coordinate axis specifies the location of an "independent item cluster." Note that although the coordinate system is graphically represented as being orthogonal Euclidean, it can be "oblique" in the probabilistic sense that $\mathrm{cov}(\Theta_i,\Theta_l) > 0$ is allowed for all $1 \le i < l \le d$, but that the matrix $\{\mathrm{cov}(\Theta_i, \Theta_l)\}$ is required to be positive definite.

Second, approximate simple structure is said to exist for a test of dimension $d \ge k$ if there exists, within the d-dimensional latent space, a k-dimensional latent coordinate system $\boldsymbol{\Theta}$ such that all items lie in narrow sectors surrounding the coordinate axes. In this case, there are k dominant dimensions. For a rigorous quantification of approximate simple structure, see Zhang & Stout (1996a). If a coordinate system not placing the items on or close to coordinate axes is used, then the k narrow sectors of items that constitute the dominant dimen-

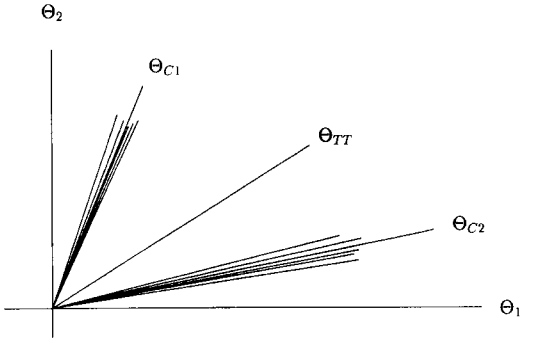
Figure 1
Geometric Representation of a Reading Comprehension Item (Item 1) and a Mathematics Story Problem Item (Item 2)



sions of the approximate simple structure will be interior to the coordinate axes (see Figure 2).

Just as each item is represented by a vector in its direction of maximum discrimination, item clusters and the entire test can also be represented in this manner. In Figure 2, Θ_{C1}, Θ_{C2}, and Θ_{TT} represent the unidimensional latent variables best measured by the two cluster scores and the total test score, respectively. The term "direction of best measurement" will be used interchangeably with "best measured unidimensional latent variable" (embedded in the $\boldsymbol{\Theta}$ space) in accordance with the geometric viewpoint.

Figure 2
A Test Demonstrating Approximate Simple Structure



The Geometry of Item Pair Conditional Covariances

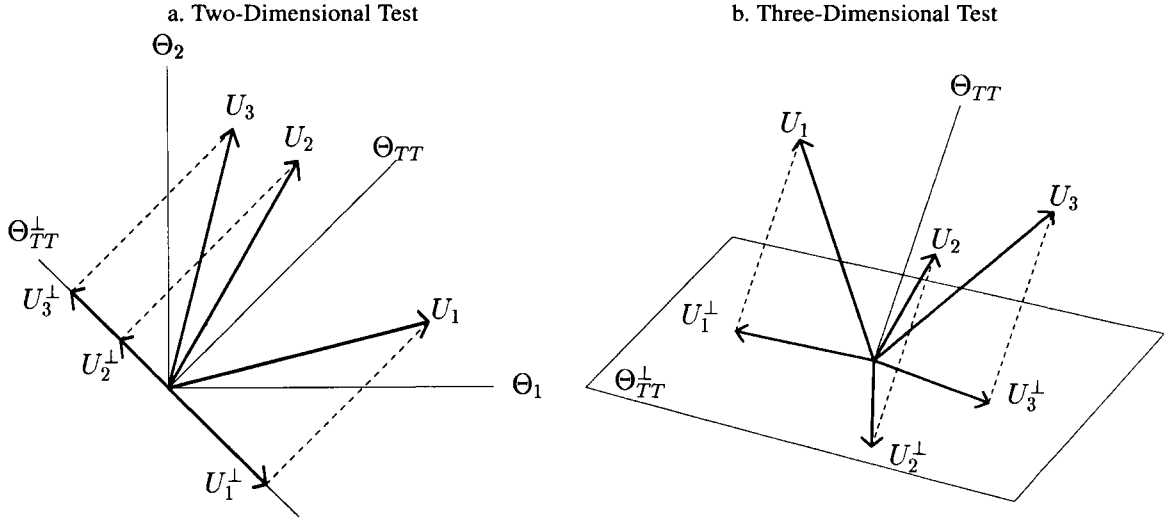
When a set of items follows a two-dimensional model, if a pair of item vectors lie on the same side of the latent conditioning variable's direction of best measurement, then the conditional covariance will be positive. If the item vectors lie on opposite sides, the conditional covariance will be negative; if either item vector lies near the conditioning variable's direction, the conditional covariance will be near 0. In Figure 2, this means that selecting one item from Cluster 1 and the other from Cluster 2 (i.e., selection from the items clustered around Θ_{C1} and Θ_{C2}) and conditioning on Θ_{TT} will result in a negative conditional covariance. Selecting both items from Cluster 1 and conditioning on Θ_{TT} will result in a positive conditional covari-

ance, and selecting both items from Cluster 1 and conditioning on Θ_{C1} will result in a conditional covariance near 0. Understanding these conditional covariance behaviors is central to understanding the three procedures described here—HCA/CCPROX, DIMTEST, and DETECT.

This reasoning has been rigorously generalized to higher dimensions (see Zhang & Stout, 1996b). For a d-dimensional test, this generalization involves expanding Reckase's geometric representation by adding the $(d-1)$-dimensional hyperplane $\Theta_{TT}^{\perp}$ (i.e., a plane for $d = 3$ and a line for $d = 2$) that is orthogonal to Θ_{TT} and passes through the origin, and projecting each item (recalling that items are represented by vectors) U_i onto this hyperplane as $U_i^{\perp}$. (Here $\perp$ denotes orthogonality and usually suggests an orthogonal projection.) This situation is illustrated for a two-dimensional test in Figure 3a and for a three-dimensional test in Figure 3b.

Zhang & Stout (1996b) showed that, for items that follow an arbitrary nonparametric compensatory multidimensional latent model (i.e., the probability of success for each item is an arbitrarily increasing function of a linear combination of the components of $\boldsymbol{\Theta}$), the deciding geometrical relationship for determining the sign of $\mathrm{cov}(U_i, U_l \mid \Theta_{TT})$ is the magnitude of the angle $\alpha_{i,l}$ between the vectors $\mathbf{U}_i^{\perp}$ and $\mathbf{U}_l^{\perp}$. For mathematical convenience, select the latent variable coordinate axes $\Theta_1, \ldots, \Theta_N$ to be uncorrelated for the population. Then $0 \le \alpha_{i,l} < \pi/2$ and the conditional covariance will be positive; if $\pi/2 < \alpha_{i,l} \le \pi$, then the conditional covariance will be negative. This last fact is of seminal importance for understanding the three procedures.

Figure 3
Expanded Geometric Representation of Three Items

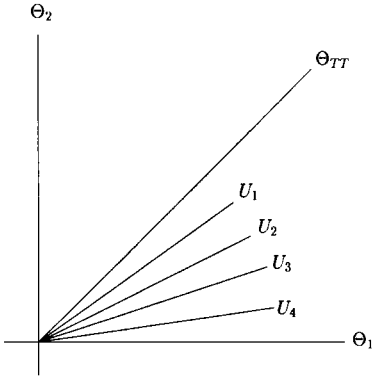


For the items in Figure 3a, note that $\mathrm{cov}(U_2, U_3 \mid \Theta_{TT})$ is positive because $U_2^{\perp}$ and $U_3^{\perp}$ face in the same direction on $\Theta_{TT}^{\perp}$; thus, $\alpha_{2,3} = 0$. However, both of the conditional covariances formed with U_1 are negative because the corresponding projections face in opposite directions on $\Theta_{TT}^{\perp}$; thus, both $\alpha_{1,2}$ and $\alpha_{2,3}$ equal π. Likewise in Figure 3b, the item pairs (1,2) and (1,3) would have negative conditional covariances, and the pair (2,3) would result in a positive value.

A theoretically more difficult problem than the behavior of the sign of $\mathrm{cov}(U_i, U_l \mid \Theta_{TT})$ is its magnitude. This problem is especially important given its direct relationship to the HCA/CCPROX procedure and the DIMTEST and DETECT statistics. Consider two items i and l whose projections on the $\Theta_{TT}^{\perp}$ hyperplane have a fixed angle $0 < \alpha_{i,l} < \pi/2$ and whose discrimination vectors have fixed magnitudes. Then, provided that $\alpha_{i,l}$ is held fixed, Zhang & Stout (1996b) strongly conjecture for these two items that $\mathrm{cov}(U_i, U_l \mid \Theta_{TT})$ is an

increasing function of each of the angles $\angle(\Theta_{TT},U_i)$ and $\angle(\Theta_{TT},U_l)$. Dynamically, this means that as item i "moves" in the (U_i, Θ_{TT}) hyperplane away from Θ_{TT} that $\text{cov}(U_i,U_l \mid \Theta_{TT})$ will increase. To illustrate this for the two-dimensional case, for the items in Figure 4, $\text{cov}(U_2,U_1 \mid \Theta_{TT}) < \text{cov}(U_2,U_3 \mid \Theta_{TT}) < \text{cov}(U_2,U_4 \mid \Theta_{TT})$. Additionally, Zhang & Stout (1996b) proved that if item i is held fixed in the latent $\boldsymbol{\Theta}$ space and item l is rotated in a cone around the Θ_{TT} axis [this cone is defined by $\angle(\Theta_{TT},U_l)$ being held constant], then if $\alpha_{i,l}$ decreases, then $\text{cov}(U_i,U_l \mid \Theta_{TT})$ will increase.

Figure 4
Directions of Best Measurement of Four Items on a Test, and for the Entire Test



Thus, $\text{cov}(U_i,U_l \mid \Theta_{TT})$ is increased by decreasing the $\alpha_{i,l}$ angle, and seems to be increased by increasing the angle either item makes with the Θ_{TT} axis. Additionally, the conditional covariance is also increased by increasing the magnitude of the items' discrimination vectors. These properties are used extensively in interpreting the results and performance of the three procedures described here. One difficulty in applying these results to statistical procedures is that when conditioning on the NC score on the remaining items, instead of on the unobservable ENC, the conditional covariance has a positive bias when unidimensionality holds (see Junker, 1993; Kim, 1994).

The LSAT

The three dimensionality assessment procedures—HCA/CCPROX, DIMTEST, and DETECT—were applied to three administrations (December 1991, June 1992, and October 1992) of the Law School Admission Test (LSAT). (For more details about the LSAT, see the 1991 Law School Admission Council report on the LSAT.) The LSAT is comprised of three item types that are grouped into sections: logical reasoning (LR), analytical reasoning (AR), and reading comprehension (RC). The operational portion of a single test consists of one AR section, one RC section, and two LR sections. For the present purposes, the two LR sections were analyzed together as a single subtest. The LR sections are comprised mostly of stand-alone items along with two or three two-item testlets per section. The AR and RC sections are each comprised of four passage-based item sets with each passage having from 5 to 8 items associated with it. Table 1 shows the number of items in each of the LSAT sections and the number of items for each passage in the AR and RC sections for the three administrations.

The LSAT dimensionality analyses summarized below appear in their entirety in Douglas, Kim, Roussos, Stout, & Zhang, 1997. Previous LSAT dimensionality analyses have been conducted with parametric dimensionality assessment tools. Camilli, Wang, & Fesq (1995) performed linear factor analyses on tetrachoric correlation matrices; Ackerman (1994) and DeChamplain (1994, 1995b) performed nonlinear factor analyses; and Reese (1995a, 1995b) estimated Yen's (1984) Q_3 statistic for all item pairs.

Table 1
Number of Items in the LR, AR, and RC Sections and in the AR and RC Passages for the December 1991, June 1992, and October 1992 LSAT Administrations

Test and Section	Administration		
	December 1991	June 1992	October 1992
LR Section			
1	25	25	25
2	24	25	26
Total	49	50	51
AR Section Passage			
1	7	6	6
2	6	5	6
3	6	6	7
4	5	7	5
Total	24	24	24
RC Section Passage			
1	7	8	6
2	8	7	6
3	5	6	8
4	8	6	7
Total	28	27	27
LSAT Total	101	101	102

HCA/CCPROX

Method

HCA/CCPROX stands for agglomerative hierarchical cluster analysis (HCA) using the unweighted pair-group method of averages with Roussos' (1995b) proximity measure CCPROX based on the item pair covariances conditioned on the remaining items. (For a general reference on clustering, see Jain & Dubes, 1988.) In agglomerative HCA, each item is initially considered to be a separate cluster. At each step, the procedure then combines two clusters from the previous step, based on the proximity between each possible pairing of clusters. The smaller the proximity measure, the more similar the two clusters are judged to be. The final clustering produced by an HCA analysis places all of the items in a single cluster. An HCA thus produces a number of clusterings equal to the number of items. For a test demonstrating approximate simple structure with k distinct item clusters, the HCA/CCPROX clustering should reproduce that structure as one of its clusterings.

For any cluster analysis procedure, the proximity measure used is very important because it is used to determine which clusters should be joined together. Because of the close connection between the behavior of item pair conditional covariances and the multidimensional geometry of test items, it seemed quite reasonable to use the item pair conditional covariances as the basis for defining a proximity measure intended to be sensitive to the dimensional distinctiveness (i.e., differences in direction of best measurement) of test items. In a test demonstrating approximate simple structure, items that belong in the same cluster will have a positive conditional covariance and those in different clusters will have negative conditional covariances. Thus, a sign change was necessary so that the proximity measure clustered the items appropriately; also, a positive constant was added so that the resulting proximity measure was non-negative.

In an extensive study, Roussos & Stout (1996) found two equally effective proximity measures based on the conditional covariance. The one used in the HCA/CCPROX procedure is

$$p_{\text{ccov}}(U_i, U_l) = \frac{-1}{\sum N_k} \sum_{k=0}^{n-2} N_k \widehat{\text{cov}}\left(U_i, U_l \middle| S_{i,l} = k\right) + \text{constant}, \tag{5}$$

where

$S_{i,l}$ is the examinee's NC score on the remaining $n - 2$ test items,

N_k is the number of examinees with $S_{i,l} = k$, and

$\widehat{\text{cov}}$ denotes the standard maximum likelihood estimate of the covariance.

The constant is added to guarantee that $p_{\text{ccov}} \geq 0$. The other equally effective proximity measure simply replaces the conditional covariance with the conditional correlation. All other proximity measures examined were significantly less effective (Roussos & Stout, 1996).

Although the proximity measure defines the proximity of two individual items, it does not provide the proximity between clusters. Among the methods of determining the proximity of clusters tested by Roussos & Stout (1996), the unweighted pair-group method of averages (UPGMA) was judged to be the most effective. The UPGMA proximity for two clusters is simply the average of all the proximities in which one item is selected from each of the two clusters. [The HCA data analysis of the LSAT in Douglas et al. (1997) was completed before that of Roussos & Stout (1996). Some of the data analyses concerning the LSAT and HCA presented here were actually conducted using the PROX proximity measure, which is both heuristically and empirically very similar to CCPROX; see Roussos (1995a) for a comparison of CCPROX and PROX.]

HCA/CCPROX Analyses of the LSAT

Ideally, a widely used standardized test will have undergone an extensive design phase, usually involving a table of specifications that the test designers and item writers attempt to follow rigorously. Thus, an examination of dimensionality often begins with a nonstatistical expert opinion assessment of the test's potential dimensionality, based on its content.

For example, the procedure used here to assess the LSAT began with the hypothesis that the AR, RC, and LR sections were dimensionally distinct, and that there would be a lesser dimensionality effect due to the passage-based construction for both the AR and RC sections. Unfortunately, sometimes the experts will miss a unifying psychological factor, or will miss dimensionally separating psychological factors. A statistical procedure such as HCA/CCPROX can thus be used as an automated second opinion to suggest various potentially dimensionally related item clusters and to help confirm those formed by expert opinion.

Figure 5 illustrates the HCA/CCPROX analysis for a random subsample of 6,000 examinees for the 24-item December 1991 AR section. Each of the 24 columns is a successive clustering of items with different clusters separated by **. For example, Level 20 shows four clusters: Cluster 1 = Items 14–19, Cluster 2 = Items 1–7, Cluster 3 = Items 8–13, and Cluster 4 = Items 20–24. In this case, at every level of the hierarchy, the clustering joined the items together according to the passage they followed, until at Level 20, the four clusters exactly corresponded to the four passages. Thus, HCA/CCPROX confirmed the previously formed expert opinion. It does, however, suggest certain other sets of items that may be of interest: Because Items 14 and 17 were the first item pair to join as a cluster, indicating that they have a large conditional covariance, Zhang & Stout's (1996) geometric interpretation of conditional covariance suggests that Items 14 and 17 were possibly the two items in a dimensionally similar cluster that differed most in the direction of best measurement from Θ_{TT}. Items 8 and 19, however, which were the last items to join their clusters, may be those items whose direction of best measurement was nearest Θ_{TT} in their respective clusters. Thus, when re-examining the cluster containing Items 14, 15, 16, 17, 18, and 19 from a cognitive or substantive perspective, Items 14 and 17 could be the most distinctive as compared to the remainder of the test, and Item 19 could be the least distinctive.

Note that although the cluster analysis suggested the possibility of additional dimensional relationships, there may be nothing dimensionally interesting in these data. In an HCA, two of the final four clusters must be selected to merge in order to form three clusters, and there must always be a first two-item cluster whether there is any actual item pair dimensional similarity or not. HCA/CCPROX does not conduct a statis-

Figure 5
HCA/CCPROX Analysis Output for a Random Sample of 6,000 Examinees for the December 1991 LSAT AR Section

Level of hierarchical cluster:

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14	14
**	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17	17
17	**	**	**	**	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15	15
**	15	15	15	15	**	**	**	**	**	**	**	**	**	**	16	16	16	16	16	16	16	16	16
15	**	**	**	**	16	16	16	16	16	16	16	16	16	16	18	18	18	18	18	18	18	18	18
**	16	16	16	16	**	**	**	**	**	**	**	**	18	18	**	**	**	**	**	19	19	19	19
16	**	**	**	**	18	18	18	18	18	18	18	18	**	**	19	19	19	19	19	**	**	05	05
**	18	18	18	18	**	**	**	**	**	**	**	**	19	19	**	**	**	**	**	05	05	06	06
18	**	**	**	**	19	19	19	19	19	19	19	19	**	**	05	05	05	05	05	06	06	03	03
**	19	19	19	19	**	**	**	**	**	**	**	**	05	05	06	06	06	06	06	03	03	04	04
19	**	**	**	**	05	05	05	05	05	05	05	05	06	06	03	03	03	03	03	04	04	07	07
**	05	05	05	05	06	06	06	06	06	06	06	06	03	03	04	04	04	04	04	07	07	01	01
05	**	06	06	06	**	**	**	**	**	**	**	03	04	04	07	07	07	07	07	01	01	02	02
**	06	**	**	**	03	03	03	03	03	03	03	04	07	07	**	01	01	01	01	02	02	12	12
06	**	03	03	03	04	04	04	04	04	04	04	07	**	**	01	02	02	02	02	**	12	13	13
**	03	**	04	04	**	**	07	07	07	07	07	**	01	01	02	**	**	**	**	12	13	11	11
03	**	04	**	**	07	07	**	**	**	**	**	01	02	02	**	12	12	12	12	13	11	09	09
**	04	**	07	07	**	**	01	01	01	01	01	02	**	**	12	13	13	13	13	11	09	10	10
04	**	07	**	**	01	01	**	**	**	02	02	**	12	12	13	11	11	11	11	09	10	08	08
**	07	**	01	01	**	**	02	02	02	**	**	12	13	13	11	09	09	09	09	10	08	**	21
07	**	01	**	**	02	02	**	**	**	12	12	13	11	11	09	10	10	10	10	08	**	21	22
**	01	**	02	02	**	**	12	12	12	13	13	11	**	09	10	**	**	**	08	**	21	22	20
01	**	02	**	**	12	12	13	13	13	11	11	**	09	10	**	08	08	08	**	21	22	20	23
**	02	**	12	12	13	13	11	11	11	**	**	09	10	**	08	**	**	**	21	22	20	23	24
02	**	12	**	13	**	11	**	**	**	09	09	10	**	08	**	21	21	21	22	20	23	24	
**	12	**	13	**	11	**	09	09	09	**	10	**	08	**	21	22	22	22	20	23	24		
12	**	13	**	11	**	09	**	**	**	10	**	08	**	21	22	**	**	20	23	24			
**	13	**	11	**	09	**	10	10	10	**	08	**	21	22	**	20	20	23	24				
13	**	11	**	09	**	10	**	**	**	08	**	21	22	**	20	23	23	24					
**	11	**	09	**	10	**	08	08	08	**	21	22	**	20	23	**	24						
11	**	09	**	10	**	08	**	**	**	21	22	**	20	23	**	24							
**	09	**	10	**	08	**	21	21	21	22	**	20	23	**	24								
09	**	10	**	08	**	21	**	22	22	**	20	23	**	24									
**	10	**	08	**	21	**	22	**	**	20	23	**	24										
10	**	08	**	21	**	22	**	20	20	23	**	24											
**	08	**	21	**	22	**	20	**	23	**	24												
08	**	21	**	22	**	20	**	23	**	24													
**	21	**	22	**	20	**	23	**	24														
21	**	22	**	20	**	23	**	24															
**	22	**	20	**	23	**	24																
22	**	20	**	23	**	24																	
**	20	**	23	**	24																		
20	**	23	**	24																			
**	23	**	24																				
23	**	24																					
**	24																						

tical test or provide a measure of the amount of multidimensionality present in the item clusters. Thus, at this stage of the dimensionality analysis it is unknown whether the four passage-based clusters were only slightly multidimensional or were highly multidimensional. An inference could be made, however, that the test is not unidimensional based on the very low probability of the four-cluster solution exactly corresponding to the four passages occurring if the exam were indeed unidimensional: Under the hypothesis of unidimensionality, the probability of this occurring is less than 10^{-10}, which provides strong evidence of a passage-based dimensionality effect. The HCA/CCPROX analysis of the other five AR and RC sections from the three LSAT administrations provided similar results—in each case the four-cluster solution corresponded exactly to the four passages.

To examine the LR sections, which were not passage-based, interpretations of the clusters formed would be more difficult. Part of this difficulty is that the sheer number of clusterings that would require examination could easily lead to faulty speculation; for example, in cases in which the clustering was not due to dimensional similarity any explanation would be erroneous. The possible dimensionality of the LR section and the causes of the dimensionality are better analyzed using HCA/CCPROX in conjunction with either DIMTEST or DETECT (this is discussed below).

DIMTEST

Method

A single run of the DIMTEST procedure (Nandakumar & Stout, 1993; Stout, 1987) assesses the conditional covariance relationship between two clusters of items on a test. The first cluster is called the assessment subtest (AT1) and is the set of items whose dimensionality is compared to the remaining items. The second cluster is called the partitioning subtest (PT) and is the set of items on which the examinees are segregated based on their score on the PT items. The intuitive idea is that examinees with the same PT score will produce approximate WLI if and only if unidimensionality holds. If PT and AT1 were selected for a test without any conscious effort to assess unidimensionality, as will be clear from the definition of the DIMTEST statistic given below, the DIMTEST statistic would merely test the hypothesis

H_0': the average of $\mathrm{cov}(U_i, U_l \mid \theta_{PT})$ over all item pairs i, l in AT1, and all θ_{PT} is ≤ 0.

Figure 6a shows a multidimensional test in which the failure to reject H_0' could be expected because of the items selected for AT1. In this example, $\mathrm{cov}(U_i, U_l \mid \theta_{PT})$ will be negative for item pairs (1,3), (1,4), (2,3), and (2,4) because they lie on opposite sides of the conditioning variable's direction of best measurement, and will be positive for item pairs (1,2) and (3,4). Thus, averaging over the six item pairs should result in a value less than 0. This example shows that failure to reject H_0' says nothing about test dimensionality. However, rejecting H_0' does imply that unidimensionality must also be rejected, because unidimensionality implies that all $\mathrm{cov}(U_i, U_l \mid \theta_{PT}) = 0$.

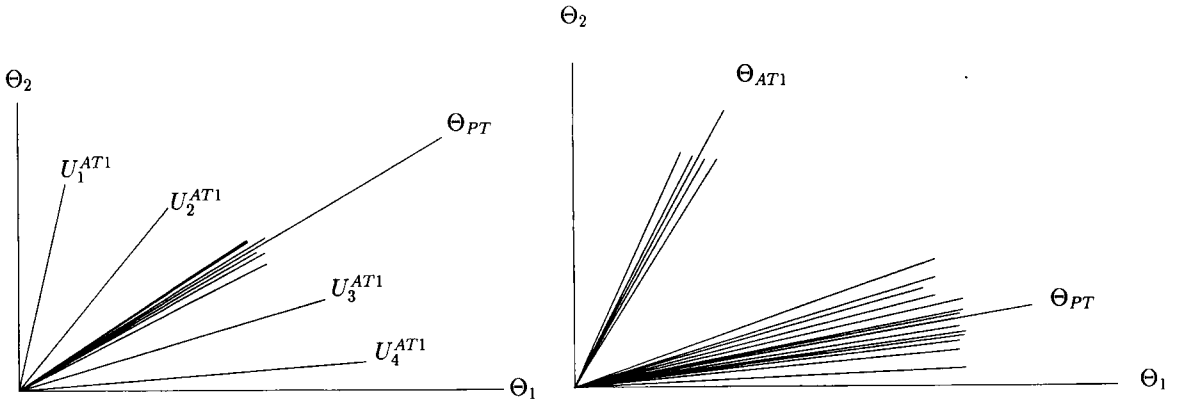
In order for DIMTEST to test with power the desired hypothesis, an informed choice of AT1 and PT must be made. For DIMTEST, the desired hypothesis is

H_0: AT1 $\cup$ PT satisfies $d = 1$

(note that AT1 $\cup$ PT may be a subtest of the original test in question). Using methods such as exploratory factor analysis (see Stout, 1987, for details; see also Hattie, Krakowski, Rogers, & Swaminathan, 1996), hierarchical cluster analysis procedures such as HCA/CCPROX, or expert opinion based on content, test structure, or cognitive considerations (see Stout, 1987), the following three principles usually lead to an effective choice of AT1 and PT.

1. The AT1 items are relatively dimensionally homogeneous. That is, geometrically the AT1 items are judged to lie in a reasonably narrow sector of the latent space. PT is not required to satisfy this requirement.

Figure 6
Poor and Good Selections of AT1 and PT for Testing Multidimensionality
a. A Poor Selection b. A Good Selection



2. The directions in which AT1 and PT best measure differ as widely as possible ("widely" means that the latent space angle between them is large).
3. The number of PT items is relatively large, and the number of AT1 items is of sufficient size to provide adequate hypothesis testing power. A suggested minimum for PT is at least 15 items; AT requires at least 3 items and not more than 1/3 of the number of PT items.

Only when it can be assumed that these principles are met can DIMTEST be considered to powerfully test H_0. If AT1 $\cup$ PT is truly unidimensional, then the vectors representing the items will coincide and each of the pairwise conditional covariances will be 0. The situation in which AT1 $\cup$ PT is multidimensional and the three principles are adhered to is shown in Figure 6b for the two-dimensional case. A similar representation would be used for $d > 2$.

In Figure 6b and in similar cases in which the three principles are adhered to, for each item pair (i, l) in AT1 and for each k,

$$\mathrm{cov}\left(U_i, U_l \middle| \Theta_{\mathrm{PT}} = k\right) > 0 \tag{6}$$

should hold, because all AT1 items lie close together and hence have separation angles $< \pi/2$ when projected on the $\Theta_{\mathrm{PT}}^{\perp}$ hyperplane [recall Zhang & Stout's (1996b) geometrically based results for the signs of item pair conditional covariances]. Summing over all such item pairs and all PT subtest ENC ks then produces the multidimensionality-indicating parameter that the DIMTEST statistic estimates. Replacing $\mathrm{cov}(U_i, U_l \,|\, \Theta_{\mathrm{PT}} = k)$ by an appropriate estimator, such as the standard maximum likelihood covariance estimator, results in the initially proposed DIMTEST statistic,

$$W' \equiv \sum_{k} \sum_{(i,l\ \mathrm{AT1\ pairs})} \widehat{\mathrm{cov}}\left(U_i^{\mathrm{AT1}}, U_l^{\mathrm{AT1}} \middle| S_{\mathrm{PT}} = k\right), \tag{7}$$

where S_{PT} is the NC score on the PT subtest. The final step in constructing the DIMTEST statistic is to compensate for the positive statistical bias caused in W'. To do this, the covariance is conditioned on the NC PT score instead of on the latent Θ_{PT}. Thus, a second assessment subtest, called AT2, is selected. AT2 is selected (1) to have the same number of items as AT1, (2) to have approximately the same item difficulty distribution as the item difficulty distribution in AT1, and (3) to have approximately the same direction of best measurement as PT. The statistic $W^{(2)}$ formed from AT2 should then reflect the amount of bias W' would show in the $d = 1$ case.

Thus, $W' - W^{(2)}$ should be an unbiased, yet powerful, index for distinguishing $d = 1$ from $d > 1$. Details of the construction and justification of the DIMTEST statistic, including obtaining its estimated standard error and a description of its asymptotic normality when $d = 1$, which allows a one-sided Z test of the hypothesis that $d = 1$, are given in Stout (1987). The resulting DIMTEST statistic is denoted by T. It is worth noting that Gao & Stout (1996) have succeeded in eliminating the need for AT2 by combining the use of Ramsay type nonparametric item response function estimation with a resampling scheme. This greatly enhances the applicability of DIMTEST and in some settings improves its statistical power.

Exploratory DIMTEST Analysis of the LSAT

The first stage of performing a statistical dimensionality assessment of a test is simply to determine whether or not it is multidimensional. If it is not, then there is no need to proceed further. Here an exploratory factor analysis was performed on a training sample of examinees in order to select optimal AT1 clusters; then DIMTEST was performed using these clusters and a cross-validation sample of examinees to compute the DIMTEST statistic. If the null hypothesis is rejected, then further exploratory analyses are necessary to determine the exact multidimensional structure.

In performing the DIMTEST preliminary analysis of the LSAT, the sample sizes for December 1991 were 6,000 for both the training and cross-validation samples; for June 1992 and October 1992 the sample sizes were 5,000. Because it was suspected that the AR, LR, and RC sections were each multidimensional, the analyses were conducted within each subtest as well as for each LSAT administration. The results are reported in Table 2 (where p is the observed level of significance.)

Table 2
Number of Items (n), T, and Observed Significance (p) From DIMTEST Exploratory Analysis

Test and	December 1991			June 1992			October 1992		
Section	n	T	p	n	T	p	n	T	p
LSAT	101	17.72	<.0001	101	22.60	<.0001	102	16.12	<.0001
AR	24	10.56	<.0001	24	14.49	<.0001	24	15.90	<.0001
LR	49	.23	.4086	50	1.61	.0540	51	2.27	.0115
RC	28	15.40	<.0001	27	13.07	<.0001	27	8.90	<.0001

Confirming the results of the HCA/CCPROX analyses, the DIMTEST results provided very strong statistical evidence ($p < .0001$) that the LSAT as a whole and the AR and RC sections were multidimensional. For LR, unidimensionality was rejected at the .05 level for the October 1992 test, was close to rejecting it for the June 1992 test ($p = .0540$), and failed to reject it for the December 1991 test ($p = .4086$).

Confirmatory DIMTEST Analysis of the LSAT

Method. Because the exploratory DIMTEST analysis indicated that multidimensionality was present in the LSAT, the next step was to test hypotheses suggested by expert opinion and the HCA/CCPROX analyses. In particular, the test developer's decision to distinguish between AR, LR, and RC items suggests testing the hypothesis that the three sections were dimensionally distinct. The hypothesis that the passage-based item clusters from the AR and RC sections were dimensionally distinct, which was strongly suggested by the HCA/CCPROX analysis, can also be formally tested.

The hypothesis that the AR, LR, and RC sections are distinct from each another involves three separate assertions: (1) that AR and LR are distinct, (2) that RC and LR are distinct, and (3) that the AR and RC are distinct. Each of these assertions must be individually tested. In each case, the AT1 items were selected from the subtest listed first, and the PT and AT2 items were selected from the other subtest. In order to distinguish the dimensional effect of various passages, whose presence was indicated by the HCA/CCPROX analysis, from the effects

of the two subtests' possibly distinct dominant dimensions, two different methods of selecting AT1 were used.

The first method was to select 8 items randomly from the entire subtest. The second method was to select 4 AT1 items in order to remove the passage-based contribution to dimensionality, by taking one item randomly from each passage in the subtest. Failure to reject unidimensionality can occur because of a poor choice of AT1, even if the subtest in question is multidimensional. Thus, each of these six hypothesis tests was conducted three times, each with a different set of AT1 items.

Results. Table 3 shows that the AR section was dimensionally distinct from both the LR and RC subtests. The two different methods of selecting AT1 gave strikingly different results for comparing the LR and RC subtests. In the case of using one item from each RC passage for AT1, the failure to reject for all nine cases seemed to very strongly indicate that for all three administrations the latent traits measured by the scores on the RC and LR subtests were highly correlated, if not identical, which is denoted by $\Theta_{RC} \approx \Theta_{LR}$, to be interpreted geometrically as these θ axes being in almost the same direction.

Table 3
Number of Rejections/Number of Hypothesis Tests From the DIMTEST Confirmatory Analysis Between LSAT Sections

Section and AT1 items	Administration		
	December 1991	June 1992	October 1992
AR vs. LR			
8 Random AR	3/3	3/3	3/3
4: One Per AR Passage	2/3	2/3	3/3
AR vs. RC			
8 Random AR	3/3	3/3	3/3
4: One Per AR Passage	3/3	3/3	3/3
RC vs. LR			
9 Random RC	3/3	3/3	3/3
4: One Per RC Passage	0/3	0/3	0/3

The rejection in the nine cases in which the AT1 items were selected randomly requires explanation. One possibility is that this was caused by a combination of the local dependence of items in AT1 selected from the same passage-based clusters, together with the over-representation of some of the clusters, resulting in a difference between Θ_{AT1} and the presumed $\Theta_{RC} \approx \Theta_{LR}$. A second possibility is that there was a difference between the direction of Θ_{RC} and Θ_{LR}, but that it was so slight that an AT1 with only four items did not allow DIMTEST adequate statistical detection power. The conclusion from this DIMTEST confirmatory analysis, that AR was distinct from LR and from RC and that the LR and RC subtests were dimensionally similar, is in agreement with the LSAT analyses of Ackerman (1994), Camilli et al. (1995), and DeChamplain (1994, 1995b).

To explore the hypothesis that each passage of the AR and RC subtests was dimensionally distinct from the other passages in the subtest, four runs of DIMTEST were made for each subtest. In each of the four runs, one of the passage-based item sets was used as AT1, and the remaining three passages comprised AT2 and PT. Subject to the difficulty distribution constraint discussed above, and the requirement that the number of items in AT1 and AT2 be the same, the assignment of items to AT2 and PT from the remaining three passages was random. For all 24 runs, the hypotheses of dimensional similarity between the passage and the remainder of the subtest were rejected with an average p value of .00002. Thus, it appears that each passage introduced its own distinct "nuisance" dimension into the AR and RC subtests. The HCA/CCPROX analysis suggested that this was apparent from the cluster analysis alone. With this strong evidence for the passage-based multidimensionality of the subtests, the next question was the amount of multidimensionality that the passages introduce. This issue was addressed by the DETECT procedure analyses presented below.

A Combined DIMTEST and HCA/CCPROX Exploratory Procedure

Method

Although HCA/CCPROX forms many clusters of possibly dimensionally distinct items, it offers no statistical test of this distinctness. However, DIMTEST is capable of testing clusters for dimensional distinctness in comparison to the remainder of the test, but offers no way of forming the clusters. A sequential method using DIMTEST and HCA/CCPROX together can alleviate each procedure's deficiencies. To avoid Type I error inflation, three random samples of the examinee population should be used: one for selecting AT1 using HCA/CCPROX, the second to conduct the DIMTEST procedure on the various HCA/CCPROX generated clusters, and the third to use DIMTEST again to confirm the dimensional structure uncovered using the second sample. One possible sequential DIMTEST-HCA/CCPROX method, which performed well in simulation studies, is as follows.

Step 1. Perform an HCA/CCPROX analysis on the test.

Step 2. Perform a DIMTEST run for each cluster of three items identified by HCA/CCPROX, using the three-item set as AT1, and the remaining items approximately split into PT and AT2 items. Note that although one-item or two-item clusters could be used, they are often psychologically uninteresting and often fail to provide adequate testing power.

Step 3. For each three-item cluster that DIMTEST shows to be dimensionally distinct from the remainder of the test's direction of best measurement, follow the cluster up the hierarchy to the next larger cluster that contains the three-item cluster. Perform a DIMTEST run on this cluster, using it as AT1, and the remaining items as PT and AT2. For example, in Figure 5 if unidimensionality was rejected for the cluster containing Items 14, 15, and 17, the next DIMTEST run done concerning these items would be done on the cluster AT1 containing Items 14, 17, 15, 16, and 18.

Step 4. If DIMTEST fails to reject the dimensional distinctiveness of PT $\cup$ AT2 from the new cluster in Step 3, then the analysis concerning that initial three-item cluster is stopped and the formerly DIMTEST-identified three-item cluster is marked as dimensionally distinct. If the new cluster is deemed dimensionally distinct, then the next cluster in the hierarchy containing it is selected as a new AT1 for a DIMTEST run. This is continued until the p value of the DIMTEST T statistic for the hypothesis test of the expanded cluster shows a noticeable increase relative to the p value of the cluster before it. When comparing the p value for the two clusters at differing levels of the hierarchy, both DIMTEST runs should be made using the same items for AT2 and for PT. For example, when comparing the p values for the cluster 14,17,15 and 14,17,15,16,18, AT2 and PT would be drawn from Items 1–13 and Items 19–24 for both DIMTEST runs (see Figure 5). The final AT1 cluster before the notable increase in the p value is marked as dimensionally distinct from the remainder of the test.

Step 5. At this point, the item clusters that were selected by DIMTEST as dimensionally distinct in Step 4 are removed from the dataset, and Steps 1–5 are repeated on the reduced dataset until no more dimensionally distinct clusters are found.

Step 6. If the clusters that are identified by Steps 1–5 contain enough items to allow use of DIMTEST, they are tested pairwise against one another (with AT1 chosen from one cluster, and AT2 and PT chosen from the other cluster) for dimensional distinctiveness, and they are also tested internally for unidimensionality, if possible. This step (when possible) allows for the actual verification or refutation of approximate simple structure.

Unfortunately, the size of the item clusters often does not allow for either part of Step 6 to be implemented. Thus, from the hypothesis testing perspective, it cannot be completely confirmed that the clusters exhibit approximate simple structure. In this case, such useful but statistically incomplete information can be augmented by the fact that it may be possible to assert with a high degree of substantive confidence (as opposed to statistical confidence) that at least some item clusters identified in the DIMTEST-HCA/CCPROX analysis exhibit approximate simple structure, by examining the cognitive content of the items within each

cluster. For example, if the sequential DIMTEST-HCA/CCPROX analysis of a passage-based test exactly identified the item clusters corresponding to the passages, it is very reasonable to conclude that the test has approximate simple structure even if the clusters are too small to perform the hypothesis testing directly.

Note that the current rule for stopping, given in Step 5, is deliberately conservative. It sometimes stops too early in the sense that adding more items would continue to preserve the dimensional homogeneity (i.e., unidimensionality) of the clusters. The reason for stopping early is illustrated by considering a dimensionally homogeneous item cluster of maximal size and considering how DIMTEST-HCA/CCPROX interacts with it.

The items in the cluster that have high discriminations tend to have closer proximities because they both have larger pairwise conditional covariances given Θ_{PT} and thus tend to join together early in the cluster analysis. However, for the lower discrimination items, the increase in the dimensional homogeneity of the cluster due to their eventual inclusion in the correct cluster may be offset by the increase in "noise" (i.e., misinformation about examinee trait level) caused by their presence. It is interesting to note that the Mokken (1971) scaling approach uses a cutoff for item inclusion in a cluster based on similar reasoning. Practically speaking, low discrimination items are simply not very informative even if they measure the correct trait. The addition of the lower discrimination items can thus increase the p value of the DIMTEST statistic that is testing the dimensional distinctiveness of the cluster relative to the remaining items on the test. This conservative stopping rule is preferred over the possibility of forming nondimensionally homogeneous (i.e., multidimensional) clusters, because in a post hoc expert opinion analysis it is often easier to recognize when two clusters are similar or when one or more items should be added to a dimensionally homogeneous cluster than when a larger cluster is actually composed of two subtly different smaller clusters.

Results of a DIMTEST-HCA/CCPROX Exploratory Analysis of the LSAT

Table 4 provides results for the AR and RC sections of the three LSAT administrations. As desired, every item was assigned to a cluster consisting only of other items of the same passage. That is, in all cases, no item was assigned to a different passage-based cluster, and none of the passage-based clusters were joined together. Furthermore, the procedure placed most of the items into their corresponding passage-based cluster, with 77% of the AR items and 89% of the RC items correctly placed into their passage-based clusters, as opposed to not having been included with any item cluster.

In addition to providing further confirmation of the passage-based approximate simple structure found by the HCA/CCPROX analysis and the DIMTEST confirmatory analysis for the AR and RC sections, the sequential DIMTEST-HCA/CCPROX greatly clarified the dimensionality assessment of the LR sections. The multitude of possible LR item clusters produced by the HCA/CCPROX analysis for each LSAT administration was refined to 4 clusters for the December administration, 7 for June, and 6 for October. These results are shown in Table 5. Examination of these clusters did produce some plausible cognitive and substantive explanations. One noticeable set of clusters consisted of (21, 22, 23, 24, 25, 46) and (48, 49) for December 1991; (25, 47, 49, 50), (19, 24, 46), and (41, 42, 44, 45) for June 1992; and (20, 24, 25, 50, 51) for October 1992. In each case, all of these items were near the end of the two timed LR sections on their respective examinations (the first LR sections all ended with Item 25 and the second LR sections ended with Items 49, 50, or 51), perhaps indicating a test speededness effect. Because the sequential method tends to be conservative, it is not unreasonable to suspect that the two clusters for the December administration and the three clusters identified for the June administration may actually belong together.

Four of the item clusters agreed with a cognitive processing analysis of the items, which identified some items as involving "additional information" (AI) and others involving "hidden assumptions" (HA) (14 AI and 12 HA for December 1991). The AI items are characterized by wording that asks the examinee to determine the piece of additional information from a multiple-choice list which, when used together with the information given in the item stem, leads to a particular logical conclusion. Cluster 1 from the Decem-

Table 4
Item Clusters Identified by the DIMTEST-HCA Exploratory Analysis for AR and RC for Three Test Administrations

Section and Cluster	December 1991	June 1992	October 1992
AR			
1	1,2,3,4,5,6,7	1,2,3,4,5,6	3,4,5,6
2	9,10,11,12,13,	7,8,9,10	7,8,9,10,11.12
3	14,15,16,17,18	12,13,14,15,16,17	13,14,15,16,17
4	20,21,22,23,24	18,19,20,21,22,23,24	20,22,23,24
RC			
1	1,2,3,4,5,6,7	1,2,3,4,5,6,7,8	1,2,3,4,5,6
2	11,12,13,14,15	10,13,14,15	7,8,9,10,11,12
3	16,17,18,20	17,18,19,21,	13,15,16,18
4	21,27,28	23,24,25,26,27	21,22,23,24,25,26,27

ber 1991 administration contained only AI items.

The HA items are characterized by wording that ask an examinee to determine what hidden assumption from a multiple-choice list was implicit in an argument that was put forth in the item stem. Cluster 3 on the December 1991 administration and six of the items in Cluster 1 on the October 1992 administration were identified as HA items. Finally, Cluster 6 for the October 1992 administration contained the only four LR items that had a historical content area, with the fifth item not having a historical content.

Table 5
Summary of the DIMTEST-HCA Exploratory Analysis Results for LR for the Three Administrations

	December 1991		June 1992		October 1992	
Cluster	Items	p	Items	p	Items	p
1	8,9,14,16	.003	7,8,9,10,12,14,15	$< 5 \times 10^{-7}$	6,10,11,14,15,34,39,33	$< 5 \times 10^{-7}$
2	21,22,23,24,25,46	.0001	16,17,20,21,22,23	$< 5 \times 10^{-7}$	17,18,19,21,22,23,46,47	$< 5 \times 10^{-7}$
3	15,37,40,41	.019	26,28,38	.003	29,32,28,37,38	.019
4	48,49	.005	31,33,29,36	.003	26,27,30,31,36,40,42	.0001
5			19,24,46	.03	20,24,25,50,51	.000001
6			25,47,49,50	.00001	4,16,41,43,49	.004
7			41,42, 44, 45.	.02		

It is apparent from the sequential analysis that there is some multidimensionality present in the LR subtest. The DIMTEST procedure does not, however, quantify the amount of dimensionality (i.e., whether the traits the sequential analysis located are minor secondary dimensions or if they have a greater effect). The estimation of the amount of dimensionality is one of the purposes for using the DETECT procedure.

DETECT

Method

The DETECT procedure was originally developed by Kim (1994) as an outgrowth of Junker & Stout's (1994) $\hat{e}$. The theoretical underpinnings and behavior of the DETECT statistic have been developed extensively by Zhang (see Zhang, 1996; Zhang & Stout, 1996a, 1996b). Whereas DIMTEST is a statistical hypothesis test and HCA/CCPROX is a sorting algorithm, DETECT is a specialized estimation procedure. The parameter being estimated by DETECT, the theoretical DETECT index $D(\mathcal{P},\Theta_{TT})$, exists for each possible partitioning, $\mathcal{P}$, of a test's items into distinct clusters and measures the amount of multidimensional approximate simple structure displayed by each such $\mathcal{P}$. Evaluating $D(\mathcal{P},\Theta_{TT})$ at the partition $\mathcal{P}$ that

maximizes the theoretical DETECT index, thus provides a theoretical index of the amount of multidimensionality present in the test and allows for the determination of whether the test exhibits approximate simple structure. "Amount of multidimensionality present" means geometrically and informally the size of the spread of the item vectors about Θ_{TT}. This is not to be confused with the number of dimensions of the latent model. In cases in which the test does exhibit approximate simple structure, the maximizing $\mathcal{P}$ partitions the test into the appropriate dimensionally distinct, internally homogeneous clusters.

For a two-dimensional test with two distinct clusters (Figure 6b), the correct partition is the one that separates the items into the two correct clusters. For this correct partitioning, $\text{cov}(U_i, U_l \mid \Theta_{TT})$ will be positive for all item pairs in the same cluster and will be negative for all item pairs that are located in distinct clusters. Using $\delta_{i,l}$ to account for this sign-cluster relationship insures that the correct partition $\mathcal{P}$ is the only partition that maximizes

$$D(\mathcal{P}, \Theta_{TT}) \stackrel{def}{=} \frac{2}{n(n-1)} \sum_{1 \le i < l \le n} \delta_{il} \mathrm{E}\left[\text{cov}\left(U_i, U_l \middle| \Theta_{TT}\right)\right], \tag{8}$$

where

$$\delta_{il} = \begin{cases} 1 & \text{if } U_i \text{ and } U_l \text{ are in the same cluster} \\ -1 & \text{otherwise} \end{cases}. \tag{9}$$

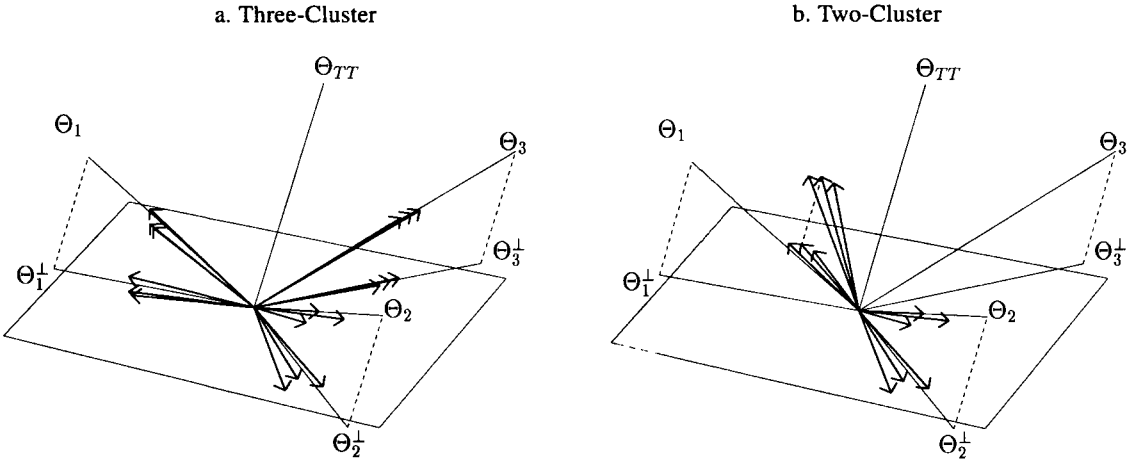
Using Zhang's geometric theory (as presented in Zhang & Stout, 1996a, 1996b), it can be proven in the case of a test of $d > 2$ that $D(\mathcal{P}, \Theta_{TT})$ will also be maximized only for the correct partition, provided that the test possesses approximate simple structure. Figures 7a and 7b illustrate two possible types of simple structure that can occur for a three-dimensional test. Figure 7a is an approximate simple structure case in which the test consists of three clusters of items, each item approximately measuring only one of the three distinct traits. Here the partition that will maximize $D(\mathcal{P}, \Theta_{TT})$ is the one dividing the test into exactly three clusters. According to the geometry of item pair conditional covariances discussed above, this is because the item projections onto the $\Theta_{TT}^{\perp}$ hyperplane produce an angular separation greater than $\pi/2$ between the clusters, thereby producing negative conditional covariances; and less than $\pi/2$ within each cluster, thereby producing positive conditional covariances, making each summand of $D(\mathcal{P}, \Theta_{TT})$ positive for the correct three-cluster partition. Thus, Θ_1, Θ_2, and Θ_3 are the axes corresponding to the three dimensions.

In Figure 7b, the test is still three-dimensional according to the definition of dimensionality based on local independence (either WLI or SLI, depending on which appears to be the most appropriate). However, it consists only of two distinct dimensionally homogeneous clusters. Although $D(\mathcal{P}, \Theta_{TT})$ is maximized for this two-cluster partition instead of for some three-cluster partition (three is the actual dimensionality), it can be argued that two is the number of dominant dimensions. For example, a reading comprehension test that is half historical context items and half scientific context items may measure three distinct traits: history, science, and reading comprehension. It is also true, and perhaps more informative, to say that it may measure two distinct dominant traits: historical context reading comprehension and scientific context reading comprehension.

DETECT's ability to count the number of dominant dimensions, as illustrated above, is linked to the occurrence of approximate simple structure. Thus, it is essential to determine whether the test does indeed have the appropriate structure. One effective method of doing this is to compare $D(\mathcal{P}, \Theta_{TT})$ to its maximum possible value

$$D^*(\mathcal{P}, \Theta_{TT}) = \frac{2}{n(n-1)} \sum_{1 \le i < l \le n} \left|\mathrm{E}\left[\text{cov}\left(U_i, U_l \middle| \Theta_{\mathrm{TT}}\right)\right]\right|. \tag{10}$$

Figure 7
Three-Dimensional Tests With Two- and Three-Cluster Approximate Simple Structure



In cases in which the test does have approximate simple structure, the ratio

$$r = \frac{D(\mathcal{P}, \Theta_{TT})}{D^*(\mathcal{P}, \Theta_{TT})} \tag{11}$$

will equal 1. The difference of a test's r from 1 is indicative of its divergence from approximate simple structure. Clearly values of r close to 1 should be viewed as still constituting approximate simple structure.

In estimating $D(\mathcal{P}, \Theta_{TT})$, it is necessary to compensate for the bias caused by conditioning on NC. Kim (1994) found the following definition to perform well:

$$D(\mathcal{P}) = \frac{2}{n(n-1)} \sum_{1 \le i < l \le n} \delta_{il} \left(\widehat{\text{cov}}_{il} - \overline{\text{cov}} \right), \tag{12}$$

where $\overline{\text{cov}}$ is the average of

$$\widehat{\text{cov}}_{il} = \frac{1}{\sum N_k} \sum_{k=0}^{n-2} N_k \widehat{\text{cov}}\left(U_i, U_l \middle| S_{i,l} = k \right) \tag{13}$$

over all pairs i,l. The definition of $\widehat{\text{cov}}_{il}$ used here is identical to that used in constructing the p_{ccov} proximity measure for the HCA/CCPROX procedure.

In addition to being an estimate of $D(\mathcal{P}, \Theta_{TT})$, which is maximized for a partition whose number of clusters is equal to the number of dominant dimensions for tests possessing approximate simple structure, the magnitude of this maximum is an indicator of the amount of multidimensionality. Based on preliminary simulation studies, maximum DETECT values near .1 or less are indicative of (possibly essential) unidimensionality (see Stout, 1990), and values greater than 1.0 indicate that sizable multidimensionality is present. Preliminary studies also indicate that values of

$$\hat{r} = \frac{D(\mathcal{P})}{D^*(\mathcal{P})} \tag{14}$$

greater than .8 are indicative of approximate simple structure. Research into the asymptotic properties of the DETECT statistic is currently in progress.

Implementation of DETECT With a Genetic Algorithm Maximization and HCA/CCPROX-Produced First Generation

In order to use DETECT, it is necessary to have an algorithm that effectively searches all partitions of the test items to find the partition $\mathcal{P}$ that maximizes $D(\mathcal{P})$, which is denoted by DETECT_{max} below. Ideally, the procedure would compute $D(\mathcal{P})$ for all possible partitions of the items into distinct clusters, but this would vastly exceed the computational ability of most computers. Therefore, some method of reducing the number of partitions to be compared is necessary. One such method [developed by Zhang (see Zhang & Stout, 1996a)] is to use HCA/CCPROX and expert opinion together to produce an initial set of partitions as a starting point for a search for the global maximizing partition $\mathcal{P}$ based on a genetic algorithm.

The first step in this genetic algorithm aided process is an HCA/CCPROX analysis of the test, and then supplementing the generated partitions with any additional partitions suggested by expert opinion. These partitions form the first "generation" of solutions. After each generation of solutions has been selected, a new generation is gained using processes analogous to the biological evolutionary concepts of "mutation" and "survival of the fittest." The new generation is initially formed by "mutating" the cluster solutions of the previous generation by randomly allowing items to be individually added or removed from the clusters, and by allowing the various partitions to randomly exchange subclusters. For each member of the new generation, DETECT estimates $D(\mathcal{P}, \Theta_{TT})$, and then those partitions with the highest values from among the newly created and the previous generation "survive" to be used to create the next generation. Additionally, some of the "less fit" partitions are allowed to contribute to the next generation for the purpose of avoiding being trapped at a local maximum. This process then repeats until a stopping criterion is satisfied (full details are available in Zhang & Stout, 1996a).

A DETECT/Genetic Algorithm Analysis of the LSAT

Consistent with the results of the DIMTEST-HCA/CCPROX sequential analysis presented in Table 5, DETECT found that the partition that maximized $D(\mathcal{P})$ (producing DETECT_{max}) corresponded exactly to the four passage-based clusters for each of the three AR subtests (Table 6). In this case, then, the genetic algorithm found as the global DETECT maximum one of the initial HCA/CCPROX generated partitions. The DETECT_{max} values for each AR administration were near 1.0, which indicated that the amount of multidimensionality was fairly strong within each AR section. Kim (1994) calculated the $\hat{r}$ value for the December AR administration as .96. The evidence is strong that the AR subtests did indeed possess passage-based approximate simple structure.

The DETECT analysis of the RC sections (also shown in Table 6) also correctly identified the four passage-based clusters without deviating from the HCA/CCPROX provided passage-based partitions for the June 1992

Table 6
Summary of the DETECT Analysis of the AR and RC Subtests

Administration	DETECT_{max}	Number of Clusters	Description of Clusters
AR Subtest			
December 1991	.9451	4	4 Passages
June 1992	1.1710	4	4 Passages
October 1992	.9770	4	4 Passages
RC Subtest			
December 1991	.7274	3	2 Science Passages Combined Plus 2 Other Passages
June 1992	.6463	4	4 Passages
October 1992	.6856	4	4 Passages

and October 1992 administrations. For the December 1991 administration, the maximum DETECT value occurred for a three-cluster partition in which two of the clusters were passage-based, and the third combined the remaining two passages. An examination of the two combined passages revealed that they were both science-based, and the other two were not. This suggests that the December 1991 administration likely had only three dominant dimensions. The observed value of $\hat{r}$ for the December 1991 administration was .93. Thus, the RC subtests also appeared to possess approximate simple structure, but the secondary dimensions were slightly weaker than for the AR subtests, with DETECT_{max} values between .64 and .73.

Unlike the AR and RC sections, recall that the HCA/CCPROX analysis of the LR section was far from conclusive. The top half of Table 7 shows that, in contrast, the DETECT solution based solely on maximizing over the first generation of HCA/CCPROX generated partitions (i.e., terminating the search for DETECT_{max} at the first generation and not using the genetic algorithm to proceed to successive generations) differed in many ways from the solutions generated by the DIMTEST-HCA/CCPROX sequential analysis. Recalling that all LR sections end either with Item 25 or approximately Item 50, the analysis did, however, replicate the earlier finding that an end-of-section effect (possibly speededness) could be present; in particular, in the top half of Table 7 the end of section predominance of the items contained in the final cluster for the December 1991 administration, the second and fourth clusters for the June 1992 administration, and the third cluster for the October 1992 administration.

The lack of agreement with the DIMTEST-HCA/CCPROX sequential analysis, and the presence of a few non-end-of-section items in the clusters purported to demonstrate speededness, is not at all surprising given the low values of DETECT_{max}, which were between .06 and .12. Similar values in simulations were indicative of situations that were very nearly unidimensional, such as when the various dimensions were very highly correlated or when the secondary dimensions had very little impact on the item responses. In

Table 7
Clusters That Maximized DETECT from First Generation DETECT and the DETECT-Genetic Algorithm for LR at Three Test Administrations

Analysis and Cluster	December 1991	June 1992	October 1992
First Generation DETECT Analysis			
1	2,14,15,18,27,29,31,32, 33,34,37,40	2,6,13,29,31,32,34,36,42 43,44,49	2,3,4,5,6,9,10,11,12,13, 15,17,18,26,27,28,29,30 31,32,33,35,36,37,38,39,44
2	17,42	17,20,21,22,23,50	7,8,19,21,34,41,42,46
3	1,6,10,11,26,28,35,36,38 39,41,43,44	26,27,28,30,33,35,38,39,41 45,46	1,14,16,20,22,23,24,25,40 43,45,47,48,49,50,51
4	3,5,7,8,9,12,13,16,20,30	25,37,47,48	
5	4,19,21,22,23,24,25,45,46 47,48,49	3,4,11,12,15,18,19,24,40	
6		1,5,7,8,9,10,14,16	
DETECT_{max}	.0682	.1177	.0903
DETECT-Genetic Algorithm			
1	2,3,5,7,8,9,10,11,12 14,15,16,17,18,27,29,30,32 33,37,38,39,40,41,42,44	3,4,6,26,27,28,29,30,31 32,33,34,35,36,37,38,39,41 42,43,44,46,48	4,8,12,14,16,17,18,19,20 21,22,23,40,41,42,43,44,46 47,48,49,50
2	4,31,34	17,20,22,23,24,25,47,49,50	2,3,5,6,7,9,13,35
3	13,19,20,21,22,23,24,25,43 45,46,47,48,49	5,7,8,9,10,11,12,14,15,16 18,19,21,45	10,11,15,26,27,28,29,30,31 32,33,34,37,38,39,45
4	1,6,26,28,35,36	1,2,13,40	24,25,51
5			1,36
DETECT_{max}	.1129	.1487	.1392

such weakly multidimensional settings, it is easy for apparently dimensionally anomalous items to be included in DETECT identified clusters. Additionally, $\hat{r}$ for the December 1991 administration was only .22, thus indicating a lack of approximate simple structure. This implied lack of simple structure, coupled with the apparent near unidimensionality of the subtest, indicates that the choice of $\mathcal{P}$ (the partition) was likely to have been overly influenced by statistical noise.

For comparison, the solutions after applying the genetic algorithm to the HCA/CCPROX found clusters are also shown in the lower half of Table 7. For all three administrations, $\text{DETECT}_{\max}$ was larger than that found by merely using the first generation of partitions selected by HCA/CCPROX. Even though the conclusions concerning the LR section did not change by using the genetic algorithm, the resulting clusters tended to agree more strongly with those found by the DIMTEST-HCA/CCPROX analysis than do those found by using the clearly suboptimal first generation DETECT results. This suggests that the genetic algorithm is a computationally efficient and often necessary method of effectively searching for the item partition maximizing $D(\mathcal{P})$.

Conclusions

HCA/CCPROX, DIMTEST, and DETECT are conditional covariance estimation-based nonparametric dimensionality assessment procedures. Each of the three procedures assesses a different aspect of test multidimensionality, and together they provide an almost complete summary of a test's dimensional characteristics, as illustrated by the preceding LSAT dimensionality case study.

The HCA/CCPROX procedure clusters test items based on their dimensional similarity as measured by Roussos' p_{ccov} proximity measure. In a test possessing approximate simple structure, one of the clusterings produced by an HCA/CCPROX analysis should group the items into the appropriate dimensionally distinct and internally homogeneous clusters. For example, in applying HCA/CCPROX to the AR and RC sections of three administrations of the LSAT, strong evidence for a passage-based dimensional structure was obtained. Even if it is unclear or doubtful that approximate simple structure holds, HCA/CCPROX can be used to generate item clusters for use as starting points for DIMTEST and DETECT analyses as well as to aid in substantive and expert opinion analyses.

When combined with expert opinion, factor analysis, or a cluster analysis procedure such as HCA/CCPROX, DIMTEST allows for testing the hypothesis that a test or subtest is unidimensional. Applied to the LSAT, DIMTEST, in conjunction with exploratory factor analysis, found that the test was multidimensional. Based on the LSAT test design intent to create three distinct item types (AR, RC, LR), DIMTEST was used in a confirmatory manner to assess the dimensional distinctiveness of the sections. The AR section was found to be dimensionally distinct from the RC section and from the LR section. However, the LR and RC sections were found to be not necessarily distinct. Working within sections, using expert opinion to select the passage-based AT1 subtests, the dimensional distinction of the various passage-based clusters was confirmed for the AR and RC subtests. When used as part of a sequential procedure with HCA/CCPROX, evidence of an end-of-section effect for LR, as well as possible cognitive- and content-based dimensions, was also obtained, in spite of the LR subtest's lack of approximate simple structure and at best weakly multidimensional structure.

The DETECT procedure, in conjunction with HCA/CCPROX and Zhang's (see Zhang & Stout, 1996a) genetic algorithm, provides an estimated measure of the degree of approximate simple structure present in a test, as well as an estimated measure of the total amount of multidimensionality present. Additionally, for a test with approximate simple structure, the partition that maximizes the $\text{DETECT}_{\max}$ index approximates the division of the items into the appropriate dimensionally distinct but internally dimensionally homogeneous clusters.

When applied to the LSAT, DETECT strongly supported the hypothesis of passage-based approximate simple structure for the AR and RC subtests, with the total amount of passage-based dimensionality being sizable ($\text{DETECT}_{\max} \geq 1$ approximately). DETECT also showed that the influence of the secondary dimen-

sions of the LR subtest was weak and that LR failed to have approximate simple structure.

In addition to the extensive statistical theory underlying these three procedures and their application to real data, such as the LSAT, extensive simulation studies have also been conducted. Studies of the effectiveness of HCA/CCPROX are found in Roussos & Stout (1996). The use of DIMTEST in conjunction with exploratory factor analysis has been reported in Stout (1987), Nandakumar (1991), Nandakumar & Stout (1993), and Hattie et al. (1996). Simulation studies of the DETECT procedure are available in Kim (1994) and Zhang & Stout (1996a).

HCA/CCPROX, DIMTEST, and DETECT are all effective at providing information about a test's dimensional structure; however, they raise many questions for future research. Although DETECT is capable of accurately identifying the items corresponding to the dominant dimensions when approximate simple structure exists, and is able to provide evidence of when a test exhibits this structure, it does not yet have a rigorous statistical asymptotic theory and is unable to provide a clear picture of the items' geometry when approximate simple structure fails to hold for tests of dimensionality greater than 2, which would also be helpful in interpreting the clusters produced by HCA/CCPROX when approximate simple structure fails to hold. The development of a nonparametric conditional covariance approach to accurately estimate the underlying geometric test multidimensionality at the item level in cases when approximate simple structure does not hold is ongoing, and includes the use of DETECT with covariance conditioning on multiple subtest scores.

DIMTEST is not limited by whether the test possesses approximate simple structure or not, but the need for an AT2 subtest to correct statistical bias limits the DIMTEST procedure in several ways, including its ability to test the dimensional homogeneity of small item clusters. A recently developed modification to remove the need for AT2 appears to be promising (see Gao & Stout, 1996).

Perhaps the most important questions raised by the capabilities of these three procedures concern the relationship between the substantive and cognitive areas a test is designed to measure and the dimensionality structure that can be found statistically. The most immediate issue is how statistical dimensionality assessment procedures can best be used by item writers and by those who interpret the test results. Clearly, one use is to detect dimensional influences not suggested by the test construction specifications, such as end-of-section speededness effects. Further, it is interesting to look for cases in which a particular category of the test's table of specifications fails to produce a statistically detectable dimensionally homogeneous cluster.

More fundamental questions include (1) Given a large examinee pool, can it be claimed that if evidence or expert opinion of a supposed cognitive or substantive difference between items cannot be verified statistically that the difference does not exist or is unimportant for test scoring or interpretive purposes? (2) If this claim can be made, how many examinees and what enhancements, if any, to these three procedures are required before the nonstatistical detection of a dimension makes it unimportant to the test practitioner or researcher? and (3) When do nonspecified but statistically detectable dimensions become important for the practitioner (e.g., speededness)? If these questions were adequately answered it would then be possible to detect cases in which test specification categories were either superfluous, or not specific enough.

References

Ackerman, T. (1994, April). *Graphical representation of multidimensional IRT analysis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans LA.

Ackerman, T. (1996). Graphical representation of multidimensional item response theory. *Applied Psychological Measurement, 20,* 311–329.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12,* 261–270.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32,* 79–96.

DeChamplain, A. (1994, January). *Assessing the dimensionality of the LSAT at the section level.* Paper presented at the University of Illinois, Department of Statistics, Champaign.

DeChamplain, A. (1995a, April). *An overview of nonlinear factor analysis and its relationship to item response theory.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

DeChamplain, A. (1995b). *Assessing the effect of multidimensionality on LSAT equating for subgroups of test takers* (Law School Admission Council Statistical Report 95-01). New Town PA: Law School Admission Council.

Douglas, J., Kim, H. R., Roussos, L., Stout, W. F., & Zhang, J. (1997). *LSAT dimensionality analysis for the December 1991, June 1992, and October 1992 administrations* (Law School Admission Council Research Report). New Town PA: Law School Admission Council.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory.* New South Wales, Australia: Center for Behavioral Studies, The University of New England.

Gao, F., & Stout, W. (1996). *Using resampling to eliminate estimation bias.* Submitted for publication.

Hattie, J., Krakowski, K., Rogers, H. J., & Swaminathan, H. (1996). An assessment of Stout's index of essential unidimensionality. *Applied Psychological Measurement, 20,* 1–14.

Hulin, C. L., Drasgow, F., & Parsons, L. K. (1983). *Item response theory.* Homewood IL: Dow Jones-Irwin.

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

Junker, B. W. (1993). Progress in characterizing strictly unidimensional IRT representations. *The Annals of Statistics, 21,* 1359–1378.

Junker, B., & Stout, W. F. (1994). Robustness of ability estimation when multiple traits are present with one trait dominant. In D. Laveault, B. Zumbo, M. Gessaroli, & M. Boss (Eds.), *Modern theories of measurement: Problems and issues* (pp. 31–61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Kim, H. R. (1994). New techniques for the dimensionality assessment of standardized test data. (Doctoral dissertation, University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 55-12B,* 5598.

Knol, D. L., & Berger, P. F. (1991). Empirical comparison between factor analysis and multidimensional item response models. *Multivariate Behavioral Research, 26,* 457–477.

Law School Admission Council. (1991). *The Law School Admission Test: Sources, contents, uses.* New Town PA: Author.

McDonald, R. P. (1981). The dimensionality of tests and items. *British Journal of Mathematical and Statistical Psychology, 34,* 100–117.

McDonald, R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6,* 379–396.

McDonald, R. P. (1994). Testing for approximate dimensionality. In D. Laveault, B. D. Zumbo, M. E. Gessaroli, & M. W. Boss (Eds.), *Modern theories in measurement: Problems and issues* (pp. 31–61). Ottawa, Canada: Edumetrics Research Group, University of Ottawa.

Mokken, R. J. (1971). *A theory and procedure of scale analysis.* New York/Berlin: De Gruyter.

Nandakumar, R. (1991). Traditional dimensionality versus essential dimensionality. *Journal of Educational Measurement, 28,* 99–117.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational and Behavioral Statistics, 18,* 41–68.

Reese, L. M. (1995a). *A comparison of local item dependence levels for the LSAT with two other tests.* Unpublished manuscript. New Town PA: Law School Admission Council.

Reese, L. M. (1995b). The impact of local dependencies on some LSAT outcomes (Law School Admission Council Statistical Report 95-02). New Town PA: Law School Admission Council.

Roussos, L. (1995a). A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and construct validity. (Doctoral dissertation, University of Illinois at Urbana-Champaign). *Dissertation Abstracts International, 57-04B,* 2956.

Roussos, L. (1995b). *Hierarchical agglomerative clustering computer program user's manual.* Urbana-Champaign: Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois.

Roussos, L., & Stout, W. F. (1996, April). *Simulation study of the effectiveness of using new proximity measures with hierarchical cluster analysis to detect simulated dimensionality structure.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait dimensionality. *Psychometrika, 52,* 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with applications to unidimensional assessment and ability estimation. *Psychometrika, 55,* 293–326.

Wang, M. (1988). *Measurement bias in the application of a unidimensional model to multidimensional item response data.* Unpublished manuscript, Educational Testing Service, Princeton NJ.

Yen, W. M. (1984). Effects of local item dependence on

the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement, 8,* 125–145.

Zhang, J. (1996). *Some fundamental issues in item response theory with applications.* Unpublished doctoral dissertation, Department of Statistics, University of Illinois at Urbana-Champaign.

Zhang, J., & Stout, W. F. (1996a, April). *A new theoretical DETECT index of dimensionality and its estimation.* Paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Zhang, J., & Stout, W. F. (1996b). *Conditional covariance structure of generalized compensatory multidimensional items with applications.* Manuscript submitted for publication.

Author's Address

Send requests for reprints or information on acquiring computer programs to execute HCA/CCPROX, DIMTEST, and DETECT to William Stout, Statistical Laboratory for Educational and Psychological Measurement, Department of Statistics, University of Illinois at Urbana-Champaign, 725 S. Wright Street, Champaign IL 61820, U.S.A. Email: stoutdist@stat.uiuc.edu. Copies of unpublished research reports and papers on these topics can be viewed and downloaded from the world wide web at http://www.stat.uiuc.edu/stoutlab.