DOCUMENT RESUME

ED 469 372 TM 034 475

AUTHOR	Douglas, Jeff; Kim, Hae-Rim; Roussos, Louis; Stout, William; Zhang, Jinming
TITLE	LSAT Dimensionality Analysis for the December 1991, June 1992, and October 1992 Administrations. Statistical Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-95-05
PUB DATE	1999-03-00
NOTE	25p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	*College Entrance Examinations; Law Schools; *Test Construction
IDENTIFIERS	DIMTEST (Computer Program); *Dimensionality (Tests); *Law School Admission Test

ABSTRACT

An extensive nonparametric dimensionality analysis of latent structure was conducted on three forms of the Law School Admission Test (LSAT) (December 1991, June 1992, and October 1992) using the DIMTEST model in confirmatory analyses and using DIMTEST, FAC, DETECT, HCA, PROX, and a genetic algorithm in exploratory analyses. Results indicate that the LSAT displays a moderate amount of multidimensionality. There appear to be two dominant dimensions. The larger seems to be created by the combination of Logical Reasoning (LR) and Reading Comprehension (RC) items. The other dimension is created by the Analytical Reasoning (AR) items. As is consistent with the apparent unidimensionality of the combination of LR and RC items, the LR and RC sections display highly correlated section scores. There is some evidence that the LR and RC sections are dimensionally distinct. DETECT indicated in two of the three administrations that the entire LSAT is three-dimensional. The LSAT also appears to have several moderately strong and easily detectable secondary dimensions, with some weaker and less easily detectable secondary dimensions associated with the LR section. It is apparent that these particular secondary dimensions introduce only weak multidimensionality relative to the entire test. (Contains 12 tables and 27 references.) (Author/SLD)

TM

LSAT Dimensionality Analysis for the December 1991, June 1992, and October 1992 Administrations

Jeff Douglas, Hae-Rim Kim,
University of Illinois at Urbana-Campaign,
Louis Roussos, Law School Admission Council,
William Stout, and Jinming Zhang,
University of Illinois at Urbana-Champaign

Law School Admission Council
Statistical Report 95-05
March 1999

LAW Services

A Publicatioo of the Law School Admissioo Couocil

The Law School Admission Council is a nonprofit corporation that provides services to the legal education community. Its members are 196 law schools in the United States and Canada.

Law School Admission Council fees, policies, and procedures relating to, but not limited to, test registration, test administration, test score reporting, misconduct and irregularities, and other matters may change without notice at any time. To remain up-to-date on Law School Admission Council policies and procedures, you may obtain a current *LSAT/LSDAS Registration and Information Book*, or you may contact our candidate service representatives.

This study is published and distributed by the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

For the purposes of this report, the dimensions of a test may be thought of as the number of statistically detectable constructs that the test is measuring. A dimensionality analysis of a test involves determining the number of dimensions being measured by the test, the nature of these dimensions, and the degree to which the dimensions are correlated. Thus, an assessment of the dimensionality structure of the Law School Admissions Test (LSAT) can a play an important role in maintaining and possibly improving the high psychometric quality of the LSAT. To this end an extensive dimensionality analysis has been conducted on the December 1991, June 1992, and October 1992 administrations of the LSAT using nonparametric tools recently developed by the authors of this report. Recent advances in the development of nonparametric tools offered the potential of powerful detection of multidimensionality while avoiding the stronger (possibly incorrect) assumptions of parametric tools.

The results of the analysis indicate that the LSAT displays a moderate amount of multidimensionality, appearing to have two dominant moderately correlated dimensions, one corresponding to the Analytical Reasoning (AR) items and one corresponding to the combined Logical Reasoning (LR) and Reading Comprehension (RC) items. These results are in agreement with the results of previously conducted parametric dimensionality analyses. The current analyses also revealed several secondary dimensions that also had been implicated by the results of previous parametric analyses. For example, our results indicate that the LR and RC sections are dimensionally distinct (though highly correlated) and that the passage-based AR and RC item sets are dimensionally distinct from each other. Finally, unlike the previous parametric analyses, our analyses also indicate that the LR items, while displaying by far the lowest level of multidimensionality of any of the item types, appear to have several weak and less easily detectable secondary dimensions. The most reliable of these weaker dimensions is an end-of-section possible speededness effect. Further analysis of these weaker dimensions is an area of future investigation because these dimensions may play a role in cognitive diagnosis or in Differential Item Functioning (DIF).

Abstract

An extensive nonparametric dimensionality analysis of latent structure was conducted on the December 1991, June 1992, and October 1992 Law School Admission Tests (LSATs) using DIMTEST in confirmatory analyses and using DIMTEST, FAC, DETECT, HCA, PROX, and a genetic algorithm in exploratory analyses. The results indicate that the LSAT displays a moderate amount of multidimensionality. The LSAT, comprised of Logical Reasoning (LR), Reading Comprehension (RC), and Analytical Reasoning (AR) sections, appears to have two dominant dimensions. The larger of the two seems to be created by the combination of LR and RC items. The other dominant dimension, the AR dimension, is created by the AR items. Consistent with the apparent unidimensionality of the combination of LR and RC items, the LR and RC sections display highly correlated section scores. Still, there is some evidence that the LR and RC sections are dimensionally distinct. The three section types taken as clusters maximized DETECT for the entire test for two of the three administrations. That is, DETECT indicated in two of three administrations that the entire LSAT is three dimensional. The LSAT also appears to have several moderately strong and easily detectable secondary dimensions. These secondary dimensions are comprised of four passage-based item sets in the AR section and four passage-based item sets in the RC section. Finally, the LSAT also appears to have several weaker and less easily detectable secondary dimensions, all associated with the LR sections. The most reliable of these weaker dimensions, an end-of-section effect, is possibly attributable to speededness associated with the LR items. The other weaker LR dimensions, while definitely present in the statistical sense, have not yet been identified substantively or cognitively with reliable, consistent interpretations. Moreover, it is clear that these particular secondary dimensions introduce only weak multidimensionality relative to the entire test.

The authorship is alphabetical. The writing of this research report was fully collaborative.

Introduction

This report presents the results of extensive nonparametric dimensionality analyses that have been conducted on three administrations of Law School Admission Test (LSAT) data: December 1991, June 1992, and October 1992. The analyses were carried out using nonparametric tools recently developed by the authors of this report. Previous LSAT dimensionality analyses have been carried out with parametric dimensionality assessment tools: Camilli, Wang, and Fesq (1995) performed linear factor analyses on tetrachoric correlation matrices; Ackerman (1994) and De Champlain (1995) performed nonlinear factor analyses; and Reese (1995a) estimated Yen's (1984) Q_3 statistic for all item pairs. Nonparametric methods were preferred for the current study because recent advances in the development of nonparametric tools offered the potential of detecting multidimensionality with the power of parametric tools and avoided the problem of falsely identifying multidimensionality due to lack of fit of an assumed parametric model.

The LSAT is comprised of three types of sections: Logical Reasoning (LR), Analytical Reasoning (AR), and Reading Comprehension (RC). A single test consists of one AR section, one RC section, two LR sections, and one unscored variable section consisting of an additional AR, RC, or LR section. For purposes of our analyses the variable section is excluded and the two LR sections are analyzed together. 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 having from five to eight items associated with them. Table 1 shows the numbers of items in each of the LSAT sections and for each passage in the AR section and RC section for the three administrations analyzed in this report.

TABLE 1

Number of items in the Logical Reasoning, Analytical Reasoning, and Reading Comprehension sections and in the Analytical Reasoning and Reading Comprehension item sets for the December 1991, June 1992, and October 1992 LSAT administrations

	LSAT Administration		
	December 1991	June 1992	October 1992
Logical Reasoning sections			
1	25	25	25
2	24	25	26
Total	49	50	51
Analytical Reasoning section item sets			
1	7	6	6
2	6	5	6
3	6	6	7
4	5	7	5
Total	24	24	24
Reading Comprehension section item sets			
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

Description of Dimensionality Assessment Procedures

A brief description of the various procedures used will help to explain the LSAT dimensionality analyses. The procedures described include DIMTEST, FAC, HCA, PROX, DETECT, and a genetic algorithm used to maximize DETECT. Additional detail is available in various manuscripts, preprints, and published papers, cited in this text.

DIMTEST

DIMTEST is a nonparametric hypothesis testing procedure developed by Stout (1987) and Nandakumar and Stout (1993) to test the null hypothesis that a specified set of items is dimensionally similar to another set of items. DIMTEST is based upon Stout's mathematically rigorous theory of essential dimensionality (Stout, 1990). Intuitively, essential unidimensionality holds for a set of items when the set of items depends on only one dominant dimension.

DIMTEST is based on detecting departures from local independence displayed by item pairs after conditioning on a unidimensional proxy for θ, the ability intended to be measured. As described in the DIMTEST user manual (Stout, Douglas, Junker, & Roussos, 1993), and in Stout (1987) and Nandakumar and Stout (1993), the general approach for one run of the DIMTEST hypothesis testing procedure can be described by the following four steps:

Step 1. After removing any items intended to be ignored in the analysis, select a group of M items from the remaining N items and call these items the AT1 subtest, where AT1 stands for Assessment Subtest One. These are the items to be tested for dimensional distinctness relative to the other $N - M$ items. AT1 can be selected based on either expert opinion or methods of exploratory data analyses, such as factor analysis or cluster analysis. By expert opinion we mean the subjective opinion of anyone who has some level of expertise in identifying psychometrically important item content by merely reading the items. In our analyses, the authors of this report performed the roles of the expert to the best of their abilities. In a more formal setting, LSAT test specialists would naturally also be consulted, but the current analyses did not involve such consultation. In general, a practitioner tries to choose AT1 so that it has the best possible chance of being dimensionally homogeneous (that is, unidimensional) with its dimension maximally distinct from the dimension, or dimensions, measured by the rest of the test.

Step 2. Select a second group of M items from the $N - M$ items that remain after the selection of AT1. This second group of M items is called the AT2 subtest. The AT2 items are chosen to be as similar as possible in difficulty level to the AT1 items while simultaneously AT2 is chosen to be dimensionally representative of the remaining $N - M$ items it comes from. Thus, AT2 should be dimensionally similar to the rest of the non-AT1 items while having a difficulty distribution similar to that of AT1. The purpose of AT2 is to eliminate a source of statistical bias, as explained below. Based on the excellent Type One hypothesis testing error rate observed in simulation studies reported in the literature, the DIMTEST AT2 selection algorithm is judged effective (see Stout, 1987; Nandakumar & Stout, 1993).

Step 3. The remaining $n = N - 2M$ items comprise the Partitioning Subtest, called PT for short. The examinees are partitioned into subgroups based on their PT scores, that is, by assigning examinees with the same number-right score, k, on PT to the same PT subgroup.

Step 4. For each PT subgroup k, a standardized difference between two variance estimates is calculated for each assessment subtest, AT1 or AT2. That is, $(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)/\hat{SE}(\hat{\sigma}_1^2 - \hat{\sigma}_2^2)$ is computed, where $\hat{SE}(\bullet)$ denotes estimated standard error of $(\bullet)$. The first of these two variance estimates $\hat{\sigma}_1^2$ is the actual observed variance of the AT1 or AT2 scores in subgroup k and the second variance estimate $\hat{\sigma}_2^2$ is the generalized binomial estimated variance of the AT1 or AT2 scores based on the assumption of unidimensionality. That is, if $\hat{p}_i$ is the proportion of examinees in cell k who answered item i correctly, then the AT1 unidimensional variance is given by $\hat{\sigma}_2^2 = \sum_i \hat{p}_i(1 - \hat{p}_i)$,

where the summation is over AT1 items. Alternatively, $\hat{\sigma}_1^2 - \hat{\sigma}_2^2$, for AT1, equals the summation of estimates of conditional item pair covariances given PT score k, where the estimate is the usual textbook covariance estimate using all examinees with PT score k (Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996). If the test is unidimensional, the standardized difference calculated for AT1 will be about the same as that for AT2. If AT1 is dimensionally distinct from PT, then the standardized difference for AT1 will be much larger than the standardized difference for AT2, recalling that AT2 is chosen to be dimensionally similar to PT. This is true because the observed variance $\hat{\sigma}_1^2$ includes the multidimensional ability contribution to examinee cell variability while $\hat{\sigma}_2^2$ does not—see Stout (1987). Thus, it is the *difference* between the AT1 and AT2 standardized differences that is the statistic of interest for each PT subgroup; and the DIMTEST statistic is based on summing this statistic of interest over all the PT subgroups. The DIMTEST statistic has been proven to be asymptotically normally distributed with mean 0 and variance = 1 when unidimensionality holds (Stout, 1987). DIMTEST rejects the hypothesis of unidimensionality when the DIMTEST statistic is greater than the upper $100(1 - \alpha)$ percentile for the standard normal distribution, α being the desired level of significance. Recent simulations by Hattie, Krakowski, Rogers, and Swaminathan (1996) led them to conclude that DIMTEST "dependably provides indications of the dimensionality, is reasonably robust, and it provides a reasonably clear and practical demarcation between one and many dimensions."

Each run of DIMTEST assesses whether the chosen AT1 is dimensionally distinct from PT with AT2 being used to eliminate statistical biasing effects. If the DIMTEST statistic were calculated without the use of the AT2 subtest, that is, based on AT1 alone, it would be biased upward because examinees are matched in subgroups using an unreliable observed score, PT. DIMTEST without AT2 could be especially biased upward if the items in AT1 are of sufficiently high discrimination and similar difficulty. This statistical bias is corrected for by the use of AT2. Nandakumar and Stout (1993) developed a modification of Stout's (1987) more conservative DIMTEST statistic that has proved more powerful in simulation studies but with a still acceptable Type One error rate. The refined DIMTEST statistic of Nandakumar and Stout (1993) is used in the applications described later in this paper. For a more thorough discussion of DIMTEST see either Stout (1987) or Nandakumar and Stout (1993).

FAC

FAC computes a tetrachoric correlation matrix for a set of dichotomous data and then performs a principal-factors factor analysis of the matrix, with maximum inter-item correlations used to estimate the communalities. Second-factor factor loadings are written to a special file for use in calculating the DIMTEST statistic. If a tetrachoric correlation is calculated as less than 0.005, then it is set equal to 0.005. If a tetrachoric correlation is calculated as greater than 0.995, then it is set equal to 0.995. No adjustments are made for guessing. If the tetrachoric correlation matrix is not positive semidefinite, the factor analysis proceeds anyway. FAC is used only as an exploratory procedure to suggest items for inclusion in AT1. As such, FAC requires no special assumptions to justify its usage, such as a linear item level factor structure or normal errors. Thus, we claim that the use of FAC with DIMTEST is still classifiable as a nonparametric approach.

HCA/PROX

HCA (hierarchical cluster analysis) is conducted using a Fortran 77 program developed by Roussos (1992) to perform agglomerative HCA (see, for example, Jain & Dubes, 1988). Agglomerative HCA starts with each object being a cluster in and of itself and successively combines pairs of clusters until all the objects join together to form one large cluster. Among many different methods for determining the proximity between clusters, the cluster distance methods known as complete link (McQuitty, 1960) and UPGMA (unweighted pair group method of averages; Sokal & Michener, 1958) have been found to give satisfactory results for our investigations. It is important to note that an HCA solution actually presents as many candidates for the best cluster partition as there are objects (items) to be clustered. Hence, for HCA to be of practical use it *must* be supplemented by some method to select the best cluster partition as the solution. In this report, DIMTEST, DETECT, and content considerations were used to supplement the HCA analyses.

Before performing a cluster analysis, a measure of proximity between all possible pairs of items is needed. The proximity should be positively associated with the degree to which the two items are measuring different dimensions. PROX, a Fortran 77 program developed by Roussos (1993), calculates one such proximity measure from Roussos, Stout, and Marden (1994), which we describe here.

In latent trait theory the concept of unidimensionality, as traditionally defined, is intrinsically linked to the concept of local independence with respect to a latent ability. The concept of local independence states that, for examinees of the same ability, the responses to different items will be statistically independent. Independence of item responses can be formulated in terms of contingency table probabilities. Because of this, a proximity measure for detecting lack of unidimensionality between two dichotomously scored items could in principle be based on the analysis of two-by-two contingency tables for the scores of examinees on the two items after examinees have been partitioned into groups of equal ability. Ability, being an unobservable latent variable, could be roughly estimated by the number-right scores of the examinees on the other items. Thus, for an M item test, the derivation of the proximity between any two items, x and y, begins with a set of M - 1 contingency tables, one for each of the possible number correct scores, k, that an examinee could have on the remaining M - 2 items. Such a contingency table is shown in Table 2.

TABLE 2
The k^{th} contingency table

	Item y score	
Item x score	1	0
1	A_k	C_k
0	B_k	D_k

In Table 2, A_k equals the number of examinees in the k^{th} contingency table who answered both items, x and y, correctly with B_k, C_k, and D_k analogously defined. Next, we let $L_k = \log(A_k D_k / B_k C_k)$, interpretable as a log odds ratio for the k^{th} contingency table. If for score k, the responses to the two items tend to be independent of each other, then

$$A_k D_k \approx B_k C_k$$

will occur and hence one expects $L_k \approx 0$.

On the other hand, strong positive dependence (high likelihood of either correct responses on both items or incorrect responses on both items) is likely to produce $L_k > 0$. Thus L_k should function well as an index of dimensional similarity in the sense that the larger the positive number, the more dimensionally similar the items. Thus it is intuitive that our proximity measure should be based upon

$$-\sum_k L_k .$$

Because one wants the proximity measure $p(x,y)$ to be approximately zero when two items are dimensionally very similar and to be nicely standardized, this suggests as reasonable the following proximity measure:

$$p(x,y) = \frac{-1}{\sqrt{n_{xy}}} \sum_k \frac{L_k}{\sqrt{Var(L_k)}} + C,$$

where n_{xy} is the number of contingency tables (number of values of k) used in the calculation of the proximity between x and y, and C is chosen so that the most similar pair of items (i.e. , the pair having the largest $\Sigma_k L_k / \sqrt{Var(L_k)}$) satisfies $p(x,y) = 0$. This is the proximity measure used in PROX. This proximity is closely related to the estimated covariance between two items conditioned on the score on the remaining items, which itself is sometimes used as a proximity measure. This estimated covariance is the numerator of DIMTEST, as remarked above. We have explored a variety of proximity measures based on the conditional covariance, and the results have shown that the PROX proximity and the conditional covariance proximity both perform very well in

simulation studies (Roussos, 1995). Our analyses presented here reflect this research in that some analyses will use the PROX proximity and others will use the conditional covariance. Clearly, the PROX proximity measure and a proximity equal to the conditional covariance will give highly correlated results.

DETECT

DETECT is useful when it is believed that a test may be composed of disjoint item clusters that are dimensionally distinct from each other and each cluster is relatively dimensionally homogeneous. Such a dimensionality structure is referred to as an (approximate) independent-clusters simple structure. DETECT is computed for a given partition of items into item clusters. For example, one might compute DETECT for the item clusters associated with the reading passage sets of the RC section of the LSAT. The intuitive principle is that the conditional covariance estimate of each item pair given the score on the remaining test should be positive or negative, subject to whether the two items in the pair belong to the same cluster or not, respectively. The index DETECT (Kim, 1994) combines non-zero second order conditional covariances of item pairs evidencing violation of unidimensionality by adding conditional covariances when two items come from the same cluster and subtracting conditional covariances when two items come from different clusters. Intuitively, the maximum value of DETECT occurs when the correct dimensionality-based cluster formation is utilized. See Stout et al. (1996) for details.

The basic objective of a DETECT analysis is to find the cluster formation that maximizes DETECT across all possible partitions into item clusters. The number of clusters for the cluster formation that maximizes DETECT is judged to be the number of dimensions present in the test, and the cluster that an item is located in corresponds to the dominant dimension the item is measuring. In addition, the magnitude of the maximum DETECT value is informative in indicating the degree of multidimensionality the test displays. Thus, DETECT helps answer the vital question of how much multidimensionality is present in a data set. In general, it is computationally prohibitive to search over all possible cluster partitions; instead, an extensive choice of cluster partitions is usually used in searching for the maximum value of DETECT. Thus, for DETECT to be effectively used, it must be assisted by an intelligent method of cluster partition selection. Both HCA and a genetic algorithm developed by Zhang (see Zhang & Stout, 1996) are useful in this regard. Simulation studies show that DETECT is highly effective at finding the correct clusters for tests having approximate independent-clusters simple structure. For further details on these studies, see Kim (1994).

Definition of DETECT

For an item pair (X_i, X_j), define a weighted sum of conditional covariances of X_i and X_j as

$$\hat{Cov}_{ij} = \frac{1}{J} \sum_{k=0}^{n-2} J_k \hat{Cov}(X_i, X_j \mid S_{ij} = k). \tag{1}$$

Here S_{ij} is the observed correct score on the $(n - 2)$ remaining items, J_k is the number of examinees with score $S_{ij} = k$, and J is the total number of examinees. The estimated covariance for the index triple (i, j, k) is computed in the usual way.

Let Ω be the set of all pairs of item indices, that is,

$$\Omega = \left\{ (i,j), 1 \le i < j \le n \right\}.$$

Note that Ω has $n(n - 1)/2$ elements. Assume that a set of clusters hypothesized to be dimensionally distinct from each other has been specified. The index DETECT is defined

$$\text{DETECT} = \frac{1}{n(n-1)/2} \sum_{(i,j) \in \Omega} (-1)^{\delta_{ij}} (\hat{Cov}_{ij} - \overline{Cov}), \tag{2}$$

$$\text{where } \delta_{ij} = \begin{cases} 0, \text{ if items } X_i \text{ and } X_j \text{ are in the same cluster,} \\ 1, \text{ otherwise.} \end{cases}$$

Here $\overline{Cov}$ is the average of $\hat{Cov}_{ij}$ over all examinee score subgroups and item pairs, and the summation in the definition (2) extends over the $n(n - 1)/2$ item pairs. The 0/1 index δ_{ij} manipulates the $(\hat{Cov}_{ij} - \overline{Cov})$ term, to be added or subtracted according to whether items X_i and X_j belong to the same cluster or not; when both items belong to the same cluster the *centered* (it is centered at $\overline{Cov}$) conditional covariance $(\hat{Cov}_{ij} - \overline{Cov})$ is added, whereas it is subtracted otherwise.

DETECT is maximized by properly assigning signs of the conditional covariances of both within- and between-cluster items through δ. Denote $DETECT_{max}$ to be the maximum DETECT calculated over all possible cluster formations. Then it becomes clear that the main objective is to find a cluster formation that maximizes, or approximately maximizes, DETECT because one suspects that the cluster structure, except for statistical error, correctly indicates the underlying multidimensional structure.

Because each conditional covariance $\hat{Cov}_{ij}$ contributes to a measure of the lack of unidimensionality resulting from violation of local independence (LI), the size of $DETECT_{max}$ can be viewed as an indicator that quantifies the amount of departure from unidimensionality. This amount of departure from unidimensionality is interpreted as the magnitude of departure from the unidimensional composite direction determined by a weighted average of all the underlying latent dimensions possibly represented by item clusters. $DETECT_{max}$ is expected to be close to zero for unidimensional data, whereas it reaches a substantially larger value for heavily multidimensional data. Based on simulation studies, Kim (1994) suggested rough categories for interpreting the amount of departure from unidimensionality that is indicated by the value of DETECT, and these categories are reproduced here in Table 3. It should be stressed that the amount of multidimensionality is distinct from the number of dimensions; a two-dimensional data set could display a large amount of multidimensionality if the two dimensions are equally well measured and are only moderately correlated whereas an eight-dimensional data set could display very weak multidimensionality if there is only one dominant dimension and/or the multiple dimensions are highly correlated. It is also important to realize that the categories of Table 3 are merely convenient choices and that for a particular application what constitutes a moderate or large amount of multidimensionality might vary considerably from that given in Table 3.

TABLE 3

A categorization of $DETECT_{max}$ *as an index of amount of multidimensionality*

$DETECT_{max}$	Multidimensionality
0.0–0.1	almost none (unidimensional)
0.1–0.5	weak
0.5–1.0	moderate
1.0–1.5	strong
1.5–and above	very strong

As mentioned earlier, in practice the search for $DETECT_{max}$ is conducted over an intelligently selected set of suspected cluster formations, rather than over all possible cluster formations, to avoid the enormous size of a combinatorial search. The next section discusses two such cluster-selection methods.

Cluster Formation

Obviously it is highly beneficial to have *a priori* a reasonable set of possibly maximizing cluster formations over which DETECT is calculated during the search for $DETECT_{max}$.

In this paper, two statistical methods for obtaining a set of cluster formations are utilized: Hierarchical Cluster Analysis (HCA) and a genetic algorithm. In using HCA with DETECT, a slightly different proximity measure than that calculated by PROX was used in the analyses presented in this paper. In generating the HCA solutions for DETECT to maximize over, the pairwise conditional covariance between items was used as the measure of the proximity between items. Also, of the many HCA options for calculating the proximity between multiple-item clusters based on the inter-item clusters, the complete link method was used. Details of HCA were given in a previous section. Details concerning the use of a genetic algorithm with DETECT are given next.

Maximizing DETECT Over All Item Cluster Partitions Using a Genetic Algorithm

Genetic algorithms are computational algorithms that use the ideas and the vocabulary from genetics and/or evolution. The purpose of a genetic algorithm approach is to optimize a function that has few nice properties (e.g., lacks derivatives). As pointed out by Michalewicsz (1994), genetic algorithms nowadays "have been quite successfully applied to optimization problems like wire routing, scheduling, adaptive control, game playing, cognitive modeling, transportation problems, traveling salesman problems, optimal control problems, database query optimization, etc."

The main idea of all genetic algorithms is that one starts with a population of possible *individuals* (i.e., *solutions*), and lets the *individuals* mutate, cross over, live and die, over successive generations until one cannot find an *individual* significantly better than the optimal individual one has gotten so far (i.e., the optimal value is "stable").

In this paper, a genetic algorithm was used to calculate the maximum DETECT value among all the k-cluster partitions of a test starting with $k = 2$ until the maximum of these maximum DETECT values was reached.

The genetic algorithm program (as developed by Zhang [Zhang & Stout, 1996]) used in the analyses reported in this paper is described as follows:

Let t = index for the generation number ($t = 1$ indicates first generation, and so on) and

$$A = (\hat{Cov}_{ij} - \overline{Cov})$$

be an $n \times n$ symmetric matrix with zero diagonal elements, where n is the number of items in the data set being analyzed. In the following four steps—which describe the genetic algorithm—each individual, parent, offspring, and so on, is a set of item clusters.

Step 1. *Obtain the initial ($t = 0$) parents for the k-cluster partitions by using HCA.* At present, we use $-A$ to obtain the item pair proximities, and use the UPGMA, complete-link, and single-link HCA methods to get three k-cluster solutions, producing three parents.

Step 2. *Produce new generation.* First, for each t^{th} generation parent randomly choose m items to mutate. Then for each parent, produce offspring by mutating the cluster membership of the parent's m randomly chosen items through all possible alternative values of cluster membership (all possible values that are different from the original values for the items). We also choose to keep clones of the parents in the offspring. Thus, there are a total of $3 \times m \times (k - 1) + 3 = N$ offspring for the new generation. In our analyses the value of m was usually taken to be between $n / 5$ and $n / 10$ (recall n = test length). Denote the new $(t + 1)^{st}$ generation of offspring by t for simplicity.

Step 3. *DETECT calculation.* Calculate the DETECT value for each offspring, and denote it as $d_i(t)$, $i = 1, 2, \ldots, N$.

Step 4. *Evaluation.* If for the new t^{th} generation $\max_{1 \le i \le N} d_i(t)$ has converged relative to the previous generations (according to a reasonable criterion that is too complex to state here), then the clustering corresponding to the maximum $d_i(t)$ value is declared to be the clustering that maximizes DETECT; its value of $d_i(t)$ is declared to be the (approximate) maximum value of DETECT. Otherwise, choose the three offspring from generation t with the largest DETECT values as parents of the next generation and return to Step 2.

Data Analysis

Three Law School Admission Tests (LSATs) administered at distinctive times (December 1991, June 1992, and October 1992) are analyzed with the tools described above. Most of the analyses include more than one procedure. The way in which the procedures were combined is explained in each case. For all of the analyses that used an initial exploratory analysis followed by some type of confirmatory analysis (for example, when cluster analysis is used in an exploratory mode to obtain AT1 for DIMTEST), two independent random samples of examinees were drawn. One, referred to herein as the *training sample*, was used in the initial exploratory analysis; the other, called the *cross-validation sample*, was used in the confirmatory analysis.

DIMTEST-FAC for Assessing $d = 1$ vs $d > 1$ (d = dimension)

An exploratory DIMTEST analysis using FAC to select the AT1 items was conducted to test for the unidimensionality of the LSAT. For the December 1991 analysis, a training sample of 6,000 randomly sampled examinees was used for the factor analysis, and a cross-validation sample of another 6,000 randomly sampled examinees was used for calculating the DIMTEST statistic. For the June 1992 and October 1992 analyses, 5,000 examinees were used in the training sample and another 5,000 were used in the cross-validation sample.

Applying DIMTEST-FAC to the entire LSAT resulted in rejections with extremely small p-values for all three administrations, strongly suggesting a lack of unidimensionality for the test taken as a whole. Thus, further exploratory DIMTEST-FAC analyses were conducted on the individual sections of the LSAT. The results for the AR and RC sections were similar as DIMTEST rejected with very small p-values for both AR and RC for all three administrations, strongly suggesting a lack of unidimensionality within each of these two sections. On the other hand, for the LR sections, the DIMTEST p-values were 0.4086, 0.0540, and 0.0115, lacking significance in two of three cases and thus showing some evidence of only one dominant dimension for the LR sections across administrations. These results replicate the dimensionality findings of De Champlain (1994). The results of all these DIMTEST-FAC analyses are presented in Table 4.

TABLE 4
Summary of DIMTEST-FAC analysis

Item Sets	LSAT Administration								
	December 1991			June 1992			October 1992		
Analysed	No. Items	T	p-value	No. Items	T	p-value	No. Items	T	p-value
LSAT	101	17.72	<0.0001	101	22.60	<0.0001	102	16.12	<0.0001
AR	24	10.56	<0.0001	24	14.49	<0.0001	24	15.90	<0.0001
LR	49	0.23	0.4086	50	1.61	0.0540	51	2.27	0.0115
RC	28	15.40	<0.0001	27	13.07	<0.0001	27	8.90	<0.0001

DIMTEST Confirmatory Analysis

Because the LSAT is divided into three different section types (LR, AR, and RC) and because two of these section types (AR and RC) are divided into passage-based item sets, the next analyses involved the use of DIMTEST to conduct confirmatory dimensionality hypothesis tests on dimensionality hypotheses naturally arising from this structure.

Between-Sections Analysis

The first set of tests was based on the hypothesis that each major section type of the LSAT is dimensionally distinct from the other two sections. To test this hypothesis each section of the LSAT was paired with each of the remaining sections, thus forming three pairs of tests: (AR,LR), (AR,RC), and (RC,LR). Then, for each pairing, AT1 was taken as a subset of the items from the first section listed above in the pairing and tested

for dimensional distinctiveness against all the items from the other section. Two methods were used to obtain AT1. In the first, AT1 was randomly sampled from the section. In the second method, AT1 was also randomly sampled but restricted to have only one item from each item set of the section that AT1 was being obtained from, thus providing protection against confounding of passage-based item sets and section effects.

It should be noted that, because the hypothesis tests were conducted at different times for the different administrations, some differences in the number of examinees used in the hypothesis testing occurred. The December 1991 administration drew a random sample of 3,000 examinees, while the June 1992 and October 1992 administrations drew a random sample of 5,000 examinees each. Because all these samples are quite large, either rerunning the December 1991 confirmatory analyses with 5,000 examinees or rerunning the 1992 analyses with 3,000 examinees was not deemed necessary as they would probably have no effect on the results.

The results of the between-sections analyses are presented in Table 5. The significance level of all the hypothesis tests was 0.05. The p-values for all of the totally random AT1s were very small—less than 0.00001. As expected, the p-values for the four-item stratified random AT1s that resulted in DIMTEST rejections were much larger because of the reduced power that comes from testing smaller AT1s. The average p-value for the four-item AT1, AR vs. LR hypothesis tests was 0.037 with five of the nine p-values at or below 0.005 and only one p-value greater than 0.10. The average p-value for the four-item AT1, AR vs. RC hypothesis tests was 0.0006 with all nine hypothesis tests rejecting $d = 1$.

TABLE 5

Summary of DIMTEST confirmatory analysis results between LSAT sections

		LSAT Administration		
Sections Tested	AT1 items	December 1991	June 1992	October 1992
Analytical Reasoning vs. Logical Reasoning	8 random Analytical Reasoning	3/3	3/3	3/3
	4: one per Analytical Reasoning passage	2/3	2/3	3/3
Analytical Reasoning vs. Reading Comprehension	8 random Analytical Reasoning	3/3	3/3	3/3
	4: one per Analytical Reasoning passage	3/3	3/3	3/3
Reading Comprehension vs. Logical Reasoning	9 random Reading Comprehension	3/3	3/3	3/3
	4: one per Reading Comprehension passage	0/3	0/3	0/3

Note. Table entries = number of rejections / number of hypothesis tests.

These results strongly confirm that the AR section is dimensionally distinct from both the LR and RC sections. It is noteworthy that even when only one item from each AR item set was used to form AT1 in testing AR against LR or RC, statistical rejection still reliably occurred. The four-item AT1 results indicate that the difference between AR and LR and between AR and RC is not simply due to the local dependence that occurs within the AR item sets. That is, the dominant dimension underlying AR is distinct from the dominant dimension underlying either LR or RC. The results for testing RC against LR show a reliable pattern of rejection when AT1 is nine randomly chosen RC items, but also show a consistent pattern of nonrejection when AT1 is restricted to being one item from each of the RC passage-based item sets. This latter result suggests that the dominant dimension underlying the RC items is the same as or very similar to the dominant dimension underlying the LR items. The statistical RC vs. LR rejections that occurred when AT1 was comprised of nine random RC items were probably due to the local dependence that occurs between items referring to the same passage, enhanced by the tendency for any particular random AT1 to have proportionally more items from some passages rather than having the items equally divided among the passages as was mandated in the four-item AT1 case. In summary, the results of this section are in agreement with the previous dimensionality analyses of Ackerman (1994), Camilli, Wang, and Fesq (1995), and De Champlain (1995), which indicated that AR is dimensionally distinct from LR and RC and that LR and RC are dimensionally very similar to each other.

Within-Section Analysis

The second set of hypothesis tests was based on the hypothesis that the AR and RC sections are each multidimensional with the items corresponding to each passage within a section forming a distinct dimension. To test this hypothesis the AR and RC item sets were analyzed separately. Within each of these sections, the items corresponding to each passage were used to form AT1 with AT2 and PT being formed from the remaining items in each section. Because each section was comprised of four item sets, four AT1s were formed for each section. The results of the within-section analyses are presented in Table 6. The *p*-values of the hypothesis tests are not displayed in the table because they were all quite small; the average *p*-value was less than 0.00002 and 21 of the 24 were less than 5×10^{-7}.

TABLE 6

Summary of DIMTEST confirmatory analysis results within Analytical Reasoning and Reading Comprehension sections

	LSAT Administration		
Section Tested	December 1991	June 1992	October 1992
Analytical Reasoning	4/4	4/4	4/4
Reading Comprehension	4/4	4/4	4/4

Notes. Table entries = number of rejections / number of hypothesis tests.
AT1 = passage-based item set within a section (four passages per section).
AT2 combined with PT = remaining items in the section.

These results strongly support the hypothesis that the item sets within the AR and RC sections form distinct dimensions. The within-section LSAT dimensionality analyses by De Champlain (1994) also indicated a lack of unidimensionality within the RC and AR sections, and confirmatory factor analyses of Camilli, Wang, and Fesq (1995) indicated that a passage-based 11-factor solution better fit their LSAT data than did a unidimensional solution or a three-dimensional item-type-based solution. Thus, both of these previously reported parametric-based dimensionality analyses are in agreement with the analyses presented here. In the case of RC, our analyses indicate that these dimensions may be simply due to the contextual information used in the passages. In the case of AR, the secondary dimensions could be due to either contextual information or simply because the same set of analytical rules applies to all the items that go with the same passage.

Of course, a DIMTEST rejection of the hypothesis of unidimensionality by itself yields no information about the amount of multidimensionality. Indeed, for large sample sizes, a test will be powerful and thus able to reject the null hypothesis even if the amount of multidimensionality is small. Thus, accurate estimation of the amount of multidimensionality is a valuable augmentation to hypothesis testing. Research is underway to evaluate the numerator of DIMTEST as such an estimator. Further, as shown in the next section, DETECT does provide an estimate of the amount of multidimensionality displayed by a partition of a test into item clusters, for example, the amount of multidimensionality in the AR section resulting from the passage-based item sets.

DETECT-HCA Exploratory Analysis

It is vital to employ a large set of meaningful cluster partitions of the test items in order to intelligently search for the maximum value of DETECT. In order to consider all possible clusterings an enormous combinatorial search would be required, which is impossible for all but very short tests. While *a priori* expert opinion may sometimes generate some clusterings to be used with DETECT, in general, an efficient and powerful exploratory method is needed to search intelligently for the maximizing partition. In this section, HCA is used to help find a cluster partition that approximately maximizes DETECT. The results of such an HCA-based DETECT analysis may then give rise through post-hoc analysis of the item wordings to further expert-opinion-based modified clusterings that can be assessed with DETECT for a possibly better simple structure solution. As mentioned earlier, the authors of this report performed the role of expert for all such analyses reported.

In the present exploratory analysis of dimensionality structure, HCA is used to provide an intelligent search for a maximum value of DETECT, thereby identifying dimensionality structure. All three LSAT item types are analyzed in this manner. By contrast, in the exploratory analysis of the dimensionality structure within the LR items—presented in the subsection *DETECT-Genetic Algorithm Exploratory Analysis of Logical Reasoning Sections*—the genetic algorithm is used to provide an intelligent search for a maximum value of DETECT, thereby identifying dimensionality structure.

Recall, when HCA is used, that the number of different clusterings formed for possible DETECT maximization is equal to the number of items analyzed. The DETECT-HCA analyses sometimes used not only the HCA clusters but also post-hoc expert-opinion clusterings derived from the HCA clusterings. Technically, the maximization referred to in the subsection *DETECT-HCA Exploratory Analysis* is over all HCA partition clusters plus possibly additional expert-opinion produced clusters when explicitly stated. HCA is preferred for use with initial exploratory DETECT analyses because HCA is much more computationally efficient than the genetic algorithm and in most cases is effective in discovering dimensionally distinct clusters that, therefore, maximize DETECT. However, for data sets for which the DETECT-HCA results are not satisfactory, further analysis using DETECT with the genetic algorithm is recommended to see if the HCA-based clusters can be improved upon. As mentioned earlier, the proximity used with HCA in the DETECT analyses was the covariance between items given score on the remaining items of the test. For all three administrations, training samples were used to generate the HCA solutions whereas distinct cross-validation samples were used to calculate DETECT. The training and cross-validation samples were 6,000 each for the December 1991 data and 5,000 each for the 1992 data sets.

The results of the DETECT-HCA analyses of the AR sections are presented in Table 7. The results were identical and striking across administrations in the sense that, for all three administrations, DETECT is maximized at HCA's four-cluster solution, which corresponded *exactly* to the four AR item sets. The range of maximum DETECT on AR is 0.9541 to 1.1710, revealing multidimensionality on the boundary between moderate and strong (c.f. Table 3). These results strongly support the suggestion that local dependence occurs between items corresponding to the same passage, leading to four distinct dimensions. This local dependence could be due to either the passage content or because a test taker's level of content mastery resulting from reading the passage creates a dependence between the items of the passage-based item set.

TABLE 7
Summary of DETECT-HCA analysis of Analytical Reasoning sections

LSAT Administration	$DETECT_{max}$	Number of Clusters	Description of Clustering
December 1991	0.9451	4	4 passage-based item sets
June 1992	1.1710	4	4 passage-based item sets
October 1992	0.9770	4	4 passage-based item sets

The results of the DETECT-HCA analyses of the RC sections are presented in Table 8. The results were very similar to those of the AR sections. In both the June and October 1992 administrations, again the four item-set clusters maximized DETECT with a moderate amount of multidimensionality according to the Table 3 categorization. The maximum DETECT value was 0.6463 for June and 0.6856 for October. HCA found this passage-based maximizing clustering for the October administration. It misidentified three items for the June administration, producing a value of DETECT slightly below 0.6463. Thus, for the June administration, we augmented the HCA-based DETECT analysis with a post-hoc expert-opinion DETECT evaluation of the four-cluster partition corresponding to the four RC item sets, which produced a new maximizing DETECT value of 0.6463. That is, the item-set partition's DETECT value was larger than any of the HCA-produced DETECT cluster values. For the December 1991 administration, the three-cluster HCA solution (a large cluster combining two passage-based item sets and two clusters corresponding to the other two passages) was found to be the clustering that maximized DETECT. The maximum DETECT was 0.7274 in this case. An inspection of the passages revealed that two science-related passages (one dealing with astronomy and one dealing with general science) combined to form the bigger cluster. Again,

contents in passages are the major factor in dividing items into dimensionally distinctive clusters. Thus, DETECT suggests three dimensions for the December 1991 administration and four dimensions for the two 1992 administrations.

TABLE 8

Summary of DETECT-HCA analysis of Reading Comprehension sections

LSAT Administration	$DETECT_{max}$	Number of Clusters	Description of Clustering
December 1991	0.7274	3	2 passage-based item sets combined plus 2 passage-based item sets
June 1992	0.6463	4	4 passage-based item sets
October 1992	0.6856	4	4 passage-based item sets

The results for the Logical Reasoning sections are presented in Table 9. These results differ from those found in the other sections in that no significantly large maximum DETECT values were found. The maximum DETECTs were 0.0682, 0.1177, and 0.0903 for December 1991, June 1992, and October 1992, with five-, six-, and three-cluster solutions from HCA, respectively. Recall from Table 3 that these levels of DETECT values are judged as evidence of either unidimensionality or very weak multidimensionality. There were no readily interpretable clusters except for what could be interpreted as end of section effects: Recall from Table 1 that the Logical Reasoning items come in two sections, one section containing 25 items and a second with 24 (December 1991), 25 (June 1992), or 26 (October 1992) items, depending on the administration. The items in Table 9 are numbered sequentially so that items 1 to 25 correspond to the first LR section for all three administrations and the remaining item numbers correspond to the second section. Thus, it can be seen that in the December 1991 and October 1992 administrations, items located mostly at the end of sections formed one cluster, showing an end-of-section effect, possibly speededness. It is noted that a cluster of items in the LR sections apparently is influenced by the end-of-section location. The small size of the maximum value of DETECT indicates that this multidimensional influence is not strong. Compared to the AR and RC sections, the LR sections appear not to have any moderate to large secondary dimensions.

Overall, these DETECT-HCA analyses exhibit strong agreement with the previously reported analyses of De Champlain (1994) and Camilli, Wang, and Fesq (1995). The within-section analyses of De Champlain (1994) indicated that the RC and AR sections exhibit a lack of unidimensionality, whereas the combined LR sections appeared to be much more unidimensional; the confirmatory factor analyses of Camilli, Wang, and Fesq (1995) indicated that the item sets corresponding to AR and RC passages cause a statistically significant amount of multidimensionality.

TABLE 9

Summary of DETECT-HCA exploratory analysis results for the Logical Reasoning sections

	LSAT Administration		
	December 1991	June 1992	October 1992
Cluster 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
Cluster 2	17,42	17,20,21,22,23,50	7,8,19,21,34,41,42,46
Cluster 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
Cluster 4	3,5,7,8,9,12,13,16,20,30	25,37,47,48	
Cluster 5	4,19,21,22,23,24,25,45 46,47,48,49	3,4,11,12,15,18,19,24,40	
Cluster 6		1,5,7,8,9,10,14,16	
DETECT value	0.0682	0.1177	0.0903

Note. Table entries: clusters that maximized DETECT-HCA.

Finally, the entire set of LSAT items for each administration was investigated as a single unit. Among clusterings from HCA and suspected clusterings using information from the previous analyses, a three-cluster solution corresponding to the three item types—AR, LR, and RC—appears to be a reasonable clustering suggesting a three-dimensional test. In fact, in the December 1991 and October 1992 administrations, the maximum DETECT values found were 0.2336 and 0.2206, respectively, which corresponded exactly to the three-cluster solution using the three item types, LR, AR, and RC. This clustering tends to confirm the presence of a three-dimensional structure. By contrast—for the June 1992 data set—a four-cluster solution (one cluster composed of the last seven AR items and the last eight RC items and three other clusters each consisting of the remaining items within each type of section) gave a maximum DETECT value of 0.2424. This suggested a four-dimensional simple structure that could be explained by three dimensions corresponding to the three item types, plus a fourth dimension corresponding to a superimposition of end-of-section and passage effects. However, it must be mentioned that several other clusterings gave values very close to the maximum DETECT found in our investigation.

DETECT-Genetic Algorithm Exploratory Analysis of Logical Reasoning Sections

Using the same training and cross-validation samples as used with DETECT-HCA, the genetic algorithm was used with DETECT to analyze the LR items. The results of the analysis are presented in Table 10.

TABLE 10

Summary of DETECT-genetic algorithm exploratory analysis results for the Logical Reasoning sections

	LSAT Administration		
	December 1991	June 1992	October 1992
Cluster 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
Cluster 2	4,31,34	17,20,22,23,24,25,47,49 50	2,3,5,6,7,9,13,35
Cluster 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
Cluster 4	1,6,26,28,35,36	1,2,13,40	24,25,51
Cluster 5			1,36
DETECT value	0.1129	0.1487	0.1392

Note. Table entries: clusters that maximized DETECT.

Comparing the bottom row of Table 10 with the bottom row of Table 9 immediately shows that the genetic algorithm was indeed able to improve upon the clusterings from HCA as indicated by the higher DETECT values obtained in the analysis. However, the values are quite small and are indicative of very weak multidimensionality, according to Table 3; again, the only clusters that were readily interpretable were those dominated by items located at the end of sections—an indication of a possible speededness dimension. Interestingly, now all three administrations produce clusters dominated by items at the ends of sections as contrasted with the DETECT-HCA analysis which failed to find an apparent end-of-section cluster for the June 1992 administration. With the AR and RC sections, the items at the end of each section also happen to be associated with the same passage. Thus, the possibility of a speededness dimension has not been discussed with respect to the RC and AR sections because such an effect, if present, would be confounded with the item-set effect also observed. One interesting pattern evident in Table 10 that was not present in Table 9 is a tendency for long strings of consecutive items (other than items that correspond to the end of a section) to appear in the same cluster. For example, in the case of December 1991, except for items 4 and 6, items 2 through 18 appear in the same cluster; except for item 43, items 37 through 44 also appear in that very same cluster. Also, in the June 1992 administration, except for item 40, items 26 through 44 appear in the same cluster. These patterns were not expected and cannot yet be adequately explained; they are unlikely to be occurring by chance. One possible explanation is that items in the same section have a tendency toward local dependence. This explanation does not fully explain the results, however, because items 2-18 and items 37-44 are in different sections, yet appear in the same cluster. Perhaps further analysis may reveal that separating each of these item sets into two different section-determined clusters will raise DETECT a little higher. Again, it is important to reiterate that the DETECT values indicate that, even if some local dependence is being detected, it is still a very weak violation of local independence. Also, because of the small DETECT values, one should not overinterpret the results. For example, the number of clusters per LR section in Table 10 is probably not a good indication of the number of dimensions.

DIMTEST-HCA Exploratory Analysis of Each LSAT Item Type

Again, the data were split into a training sample and a cross-validation sample. First, PROX was used on the training sample to compute a proximity matrix for the test items. Then, an UPGMA unweighted pair-group method of averages HCA solution was generated. The HCA solution is used to partition a test into clusters of items that potentially contribute to secondary dimensions on the test. Note that the use of PROX and UPGMA in the analyses presented here may sometimes give slightly different HCA solutions than those

obtained in the DETECT-HCA analyses presented earlier that used a different proximity measure (conditional covariance) and a different HCA method (complete link). Then, using the cross-validation sample, DIMTEST was used to test for the dimensional distinctiveness of these clusters. Two disjoint 3,000-examinee samples were extracted from the December 1991 data, and two disjoint 5,000-examinee samples were drawn from both the June 1992 and the October 1992 data for use as training and cross-validation samples.

Using DIMTEST With HCA to Identify Dimensionally Distinct Clusters

Before presenting the results of the data analysis, we review how DIMTEST is used in conjunction with HCA to identify dimensionally homogeneous clusters.

Identifying clusters at initial stage of development. Clusters that form early in the hierarchy of the HCA solution are used as trial AT1s in the DIMTEST procedure. With each such AT1 tested, PT and AT2 together consist of all the remaining items on the test. Even though the initial two-item clusters sometimes result in DIMTEST statistical rejection, clusters containing at least three items usually need to be considered. Moreover, two-item clusters seem somewhat uninteresting from the psychological perspective and should usually be ignored in statistical dimensionality analyses.

Following development of clusters up the hierarchy. Once a cluster has been identified that results in rejection of the DIMTEST statistic, the cluster should be followed up the hierarchical solution as it joins with other clusters. Each time the initial cluster joins with another cluster, the new joint cluster would be tested with DIMTEST and this new joint cluster would then be followed until it joined with another cluster, and so on. This procedure of following expanding clusters up the hierarchy should be done with every cluster that causes rejection of the DIMTEST statistic.

Stopping decision. If in following an increasing cluster up the hierarchy, continued rejection of the DIMTEST statistic is found, the question arises as to when to stop. Several plausible ways to proceed are discussed in Roussos, Stout, and Marden (1994). This paper uses the following stopping rule: proceed up the HCA solution until the DIMTEST p-value of a new joint cluster noticeably increases relative to the p-value of the preceding cluster before the new cluster was added to it. This stopping rule will usually be conservative, leading to stopping decisions that are often premature, in that adding more items would continue to preserve the dimensional homogeneity of the cluster. To understand this phenomenon, let us consider a set of items all measuring the same dimension. The items with the higher discriminations will tend to have closer proximities and thus tend to be joined together first in the cluster analysis. Hence, the increase in dimensional homogeneity that comes from more items joining a cluster will be offset by the noise resulting from the lower discrimination of these added items. The last items to join a cluster may often cause a greater increase in noise from their lower discrimination than the reduction in noise due to the correct inclusion of the additional items that are dimensionally similar to the items already selected for the cluster, thus increasing the p-value of the DIMTEST statistic testing the dimensional distinctiveness of the cluster relative to the remaining items on the test. Still, this conservative stopping rule was preferred so that the identified clusters would be as homogeneous as possible. Such homogeneous clusters will be more easily interpretable in a substantive expert-opinion post-hoc analysis. Consequently, any remaining unclassified items may be classified in the correct cluster by comparing the content of the unclassified items to the common content of items in each formed cluster.

Sequential DIMTEST-HCA analysis. In using HCA with DIMTEST to identify dimensionally distinct clusters, a sequential technique is often helpful. Some HCA clusters that cause rejection of the DIMTEST statistic may fail to cause rejection when they join with other clusters further up the hierarchy. Also, though some clusters may cause DIMTEST statistical rejection, many other clusters will not cause rejection. At such a point, it would appear that no further dimensionally distinct clusters could be identified. However, it is well known that HCA solutions tend to migrate away from the optimal solution, that is, from the most accurate portrayal of the multidimensional reality, as the solution progresses further up the levels of the hierarchy (Hubert & Baker, 1979). Although the initial clusters may be fairly accurate, more and more statistical error tends to creep in as the solution progresses up the hierarchy. Thus, after identifying all the item clusters that cause DIMTEST rejection in the initial analysis,

a useful technique is to generate a second HCA solution on only those items that did not fall into the clusters that caused DIMTEST rejection in the initial analysis. DIMTEST would then be used with this second HCA solution in the same manner as described above to try to identify more dimensionally distinct clusters. This sequential procedure can be performed repeatedly until no new clusters are identified that reject with DIMTEST or until the number of remaining items is so small that DIMTEST hypothesis testing can no longer be carried out. This sequential technique was applied in the analysis of the LSAT data below.

Note that the DIMTEST-HCA analysis ends without a firm statistical determination of overall dimensional structure, although with *highly* useful partial information about the dimensional structure. Such statistical incompleteness is sometimes offset by a fairly large degree of substantive confidence (as opposed to statistical confidence) that at least some item clusters identified in a DIMTEST-HCA analysis are indeed dimensionally distinct from one another (and perhaps even essentially unidimensional within cluster) by a substantive examination of the cognitive content of the items in the identified clusters.

DIMTEST-HCA Exploratory Analysis of the Reading Comprehension and Analytical Reasoning Sections

The DETECT-HCA results already have demonstrated that these two sections appear, in general, to be four dimensional with the four dimensions being the four-item sets in each section. In this analysis using PROX and UPGMA, for all three administrations, the HCA four-cluster solution for both the AR and RC sections was *precisely* the four item sets associated with the four passages. As indicated above, we expect the DIMTEST-HCA exploratory analysis to select clusters that occur lower in the HCA hierarchy than does the four-cluster solution.

The results of the DIMTEST-HCA exploratory analyses of the RC and AR sections are presented in Table 11. In all cases the DIMTEST p-value was less than 0.0000005. The strong agreement between these results and DETECT-HCA results serves to support the claim that the DIMTEST-HCA analysis is effective at identifying homogeneous clusters. The results also show that the DIMTEST-HCA stopping rule is indeed conservative, as indicated by some items not being classified into any clusters. However, if the statistical testing had progressed far enough up the HCA solution, the four-cluster solution would have corresponded perfectly to the four item sets in each of the two sections. On average, 77% of the RC items and 89% of the AR items were classified by the purely statistical part of the DIMTEST-HCA analyses without any post-hoc substantive investigation of the clusters.

TABLE 11

Summary of DIMTEST-HCA exploratory analysis results: Reading Comprehension and Analytical Reasoning sections

	LSAT Administration		
Section and Cluster	December 1991	June 1992	October 1992
Analytical Reasoning			
Cluster			
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
Reading Comprehension			
Cluster			
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

Note. Table entries = clusters found in the DIMTEST-HCA analysis.

DIMTEST-HCA Exploratory Analysis of the Logical Reasoning Sections

The DETECT-HCA and DETECT-genetic algorithm results indicated that LR sections appear nearly unidimensional, though clusters existed that suggested there may be end-of-section effects and perhaps other minor dimensions. Simulation studies have shown that DETECT has the greatest difficulty detecting dimensionality structure when the items are of mixed structure (each item measures more than one dimension in a manner such that dimensionally distinct clusters do not exist), the dimensions are highly correlated (0.9 or greater), and/or the number of items measuring a given dimension is small in size (Kim, 1994). Thus, the DIMTEST-HCA analysis was conducted to try to tease out any possible weaker sources of multidimensionality that would not be evident in the DETECT results. The results of the DIMTEST-HCA exploratory analysis of the LR sections are presented in Table 12. For all three administrations, clusters of items are identified that are associated with the ends of the two LR sections, thus indicating the presence of a possible speededness dimension. Indeed, post-hoc confirmatory DIMTEST analyses with all three administrations showed strong rejections of end of section and hence possible speededness item sets. For the December 1991 data, an AT1 consisting of items 21 to 25 and 45 to 49 rejected with a p-value of 0.0004. For the June 1992 data, an AT1 consisting of items 21 to 25 and 46 to 50 rejected with a p-value of 0.0004. For the October 1992 data, a much smaller AT1 was tested because with this data set too many of the most difficult items on the test were also at the end of the LR sections making it difficult to find an appropriate AT2. So, for October 1992, AT1 was set equal to items 24, 25, 49, and 51, which rejected with a p-value of 0.006.

TABLE 12

Summary of DIMTEST-HCA exploratory analysis results: Logical Reasoning sections

LSAT Administration					
December 1991		June 1992		October 1992	
Clusters	p-value	Clusters	p-value	Clusters	p-value
8,9,14,16	0.0030	7,8,9,10,12,14,15	$< 5 \times 10^{-7}$	6,10,11,14,15,33,34,39	$< 5 \times 10^{-7}$
21,22,23,24,25,46	0.0001	16,17,20,21,22,23	$< 5 \times 10^{-7}$	17,18,19,21,22,23,46,47	$< 5 \times 10^{-7}$
15,37,40,41	0.0190	26,28,38	0.00300	28,29,32,37,38	0.019000
48,49	0.0050	29,31,33,36	0.00300	26,27,30,31,36,40,42	0.000100
		19,24,46	0.03000	20,24,25,50,51	0.000001
		25,47,49,50	0.00001	4,16,41,43,49	0.004000
		41,42,44,45	0.02000		

In addition to possible speededness clusters, a number of other clusters were found, more so for June 1992 and October 1992 than for December 1991. This difference may be due to the use of data from only 3,000 LSAT examinees to analyze the December 1991 data as compared with 5,000 for the analysis of the 1992 data. However, the other analyses presented in this paper, in which 6,000 examinees were used with the December 1991 data, also showed that data set to be more nearly unidimensional than the two 1992 data sets. Some of these other non-end-of-section clusters were quite large and rejected with quite small p-values. Most notably in this regard were the clusters (7,8,9,10,12,14,15) with June 1992 and (6,10,11,14,15,33,34,39) with October 1992. A cursory inspection of these and the other non-end-of-section clusters sometimes revealed a homogeneous content area, more often some cognitive processing similarities, but most often no quickly recognizable secondary dimension. Indeed one cluster of items was determined to be an artifact of varying difficulty levels. The cluster (26,28,38) on June 1992 is comprised of the three easiest items on the test and there were no other items of approximately equal difficulty that could go into AT2 to balance them. Thus, it was concluded that this particular cluster was not an indication of multidimensionality. Other post-hoc analyses, however, were more fruitful. A cognitive processing analysis of the December 1991 item wordings indicated several major dimensions could be underlying the item responses, the two largest of which, for purposes of this report, we refer to as *additional information* (AI) items (14 of the 49 December 1991 LR items) and *hidden assumption* (HA) items (12 items). The AI items are characterized by wordings that ask the examinee to determine the piece of additional information from a multiple-choice list, that, when used together with the information given in the item stem, leads to a particular conclusion. The HA items are

characterized by wording that asks an examinee to figure out what hidden assumption from a multiple-choice list was implicit in an argument put forth in the item stem. In the December 1991 analysis, the (8,9,14,16) cluster was found to be all AI items, and the (15,37,40) cluster was all HA items. In the June 1992 analysis, of the five items (8,9,10,12,14) that could be easily labeled as either HA or AI in a cursory analysis, four were found to be AI items. And in the October 1992 analysis, six of the eight items in the cluster (6,10,11,14,15,33,34,39) were identified as HA items. One cluster in the October 1992 analysis did seem to have a homogeneous content area: In the cluster (4,16,41,43,49), all but item 49 involved the content area of history and those four items were the only items on the October 1992 LR that involved history.

There seems little doubt from these analyses that the LR sections are not strictly unidimensional in that reliable statistical rejection of the DIMTEST null hypothesis was found for all three administrations; however, these results do not reflect on the strength of the lack of unidimensionality. Indeed, as the DETECT results have shown, the multidimensionality within the LR items is very weak. The results here merely demonstrate that this weak multidimensionality is not to be equated with unidimensionality. The nature of this multidimensionality, aside from an end-of-section effect is not yet fully clear. While cognitive processes and content areas have been implicated as possible factors, further analyses are necessary to determine whether these or other factors are truly the cause of the multidimensionality.

Concluding Remarks

An extensive nonparametric dimensionality analysis has been conducted on the December 1991, June 1992, and October 1992 LSAT administrations using DIMTEST in confirmatory analyses and using DIMTEST, FAC, DETECT, HCA, PROX, and a genetic algorithm in exploratory analyses. The results indicate that the LSAT displays a moderate amount of multidimensionality with the LSAT appearing to have two moderately correlated dominant dimensions, one corresponding to the AR section and one corresponding to the combined LR and RC sections. These results are in agreement with the previously conducted parametric dimensionality analyses of Ackerman (1994), De Champlain (1995), and Camilli, Wang, and Fesq (1995), which all indicated that the LSAT dimensionality structure is primarily governed by these same two dimensions. Of these two dominant dimensions, the larger is the one that results from the combined influence of the LR and RC items. That is because these two dimensions combine to make up the vast majority of items on the LSAT (approximately 75% of the total number of items). Even though the AR section is dimensionally distinct from the LR and RC sections, it is important to point out that the AR section is highly correlated with the LR and RC sections (e.g., for the December 1991 administration, number-right score on AR had observed correlations of 0.589 and 0.487 with number-right score on LR and RC, respectively), though not as strongly correlated as LR is with RC (0.743 in the above case).

The current dimensionality analysis delved deeper into the dimensionality structure of the LSAT to reveal several secondary dimensions (weaker than the two dominant dimensions) that had previously been implicated by the results of De Champlain (1994), which showed that the AR and RC sections lacked unidimensionality, and by the results of Camilli, Wang, and Fesq (1995), which indicated that at least some (if not all) passage-based testlets produce violations of local independence. For example, our results give some evidence that the RC and LR sections are dimensionally distinct, because the three section types taken as clusters tended to maximize DETECT for the entire test. The LSAT also appears to have several other moderately strong and easily detectable secondary dimensions. These secondary dimensions are the four passage-based item sets in the AR section and the four passage-based item sets in the RC section. Finally, the LSAT also appears to have several weaker and less easily detectable (and not previously detected) secondary dimensions, all associated with the LR sections. The most reliable of these weaker dimensions is the end-of-section possible speededness effect for the LR items. The other weaker LR dimensions, although definitely present in the statistical sense, have not yet been identified substantively or cognitively with reliable consistent interpretations. Moreover, it is clear that these particular secondary dimensions introduce only weak multidimensionality relative to the entire test. Further analysis of these weaker dimensions is an area of future investigation because these dimensions may play a role in cognitive diagnosis or in causing differential item functioning (DIF).

The multidimensionality results reported herein raise a large number of questions and future research possibilities. First, do the findings replicate for more recent LSAT administrations, such as the 1994 and 1995 administrations? Are the RC and LR section types, which have quite different and very detailed item-writing specifications, really measuring essentially the same construct from the statistical data analysis perspective? What are the implications for LSAT construction, analysis, and interpretation of the very consistently demonstrated passage dimensionality effect for the AR and RC sections? Is the AR passage effect primarily content-based or due more to the common influence on all passage items of the examinees solution model for the passage, or are both influences important? Do some item set content areas have more of a dimensional influence than others? How do the variety of content choices relate to the overall construct validity of the LSAT? Is there on occasion a content effect (recall the history item cluster) on the LR? Are the author's AI and HA category dimensions replicable? Are there other statistically detectable cognitively based dimensions on the LR? How do these categories cut across and relate to the fine-grained item types of the LR item-writing specifications? How do these findings relate to the unified cognitive/psychometric model by DiBello, Stout, and Roussos (1995)? Do the various weak to low moderate secondary dimensions found have measurement implications or are they small enough to be ignored? What are the implications of the very clearly demonstrated multidimensionality of the LSAT for future computerized LSAT development and analyses? How does this research compare and contrast with the local dependence/multidimensionality research of De Champlain (1994; 1995), Reese (1995a; 1995b), and Thissen (Chen & Thissen, 1997), among others? What are the connections with Ackerman's work on geometric representation of multidimensionality at the item level? The above represent a sampling of the many questions raised by our multidimensionality analysis of the LSAT.

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.

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.*

Chen, W. -H., & Thissen, D. (1997). Local dependence indices for item pairs using item response theory. *Journal of Educational and Behavioral Statistics,* 22, 265-289.

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

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

DiBello, L. V., Stout, W. F., & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In P. D. Nichols, S. F. Chipman, & R. L. Brennan (Eds.), *Cognitively diagnostic assessment* (pp. 361-389). Hillsdale, NJ: Lawrence Erlbaum Associates.

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.

Hubert, L. J. & Baker, F. B. (1979). Identifying a migration effect in a complete-link hierarchical clustering. *Journal of Educational Statistics,* 4(1), 74-92.

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

Kim, H.R. (1994). *New techniques for the dimensionality assessment of standardized test data.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign.

McQuitty, L.L. (1960). Hierarchical linkage analysis for the isolation of types. *Educational & Psychological Measurement, 20,* 55-67.

Michalewicsz, Z. (1994). *Genetic algorithms + data structures = genetic programs.* Berlin: Springer-Verlag.

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

Reese, L. M. (1995a). *A comparison of local item dependence levels for the LSAT with two other tests.* Unpublished manuscript.

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

Roussos, L. A. (1992). *Hierarchical agglomerative clustering computer program users manual.* Unpublished manuscript, University of Illinois at Urbana-Champaign, Champaign.

Roussos, L. A. (1993). *PROX help sheet.* Unpublished manuscript, University of Illinois at Urbana-Champaign, Champaign.

Roussos, L. A. (1995). *A new dimensionality estimation tool for multiple-item tests and a new DIF analysis paradigm based on multidimensionality and construct validity.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Champaign.

Roussos, L. A., Stout, W. F., & Marden, J. I. (1994). *Analysis of the multidimensional structure of standardized tests using DIMTEST with hierarchical cluster analysis.* Unpublished manuscript.

Sokal, R. R., & Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin, 38,* 1409-1438.

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.

Stout, W. F., Douglas, J., Junker, B., & Roussos, L. A. (1993). *DIMTEST manual.* Unpublished manuscript available from W. F. Stout, University of Illinois at Urbana-Champaign, Champaign.

Stout, W. F., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance based nonparametric multidimensionality assessment. *Applied Psychological Measurement, 20,* 331-354.

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., & Stout, W. F. (1995, April). *Theoretical results concerning DETECT.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Zhang, J., & Stout, W. F. (1996, 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.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

Reproduction Basis

[X] This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

[ ] This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").