DOCUMENT RESUME

ED 468 754 TM 034 431

AUTHOR	Camilli, Gregory; Wang, Ming-mei; Fesq, Jaqueline
TITLE	The Effects of Dimensionality on True Score Conversion Tables for the Law School Admission Test. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-92-01
PUB DATE	1992-06-00
NOTE	33p.
PUB TYPE	Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	Admission (School); *College Entrance Examinations; Factor Analysis; Item Response Theory; Law Schools; Law Students; Thinking Skills; *True Scores
IDENTIFIERS	*Dimensionality (Tests); *Law School Admission Test

ABSTRACT

The Law School Admission Test (LSAT) was examined to see if the items on a form could be divided into different subgroups in which items looked statistically similar within the subgroups but statistically different between subgroups. Of such subgrouping can be detected, it is likely that the subgroups of items measure different abilities, and the test can be described as "multidimensional." The multidimensionality of six forms of the LSAT was studied using factor analysis. Two subgroups of items, or "factors," were found for each of the six forms. The LSAT thus appears to measure two different reasoning abilities, inductive and deductive. The effect of dimensionality on equating was also examined by calibrating, with item response theory (IRT) methods, all items on a form to obtain a set of estimated item parameters (Set 1). The test was then divided into two homogeneous subgroups of items, each having been determined to represent a different ability. Items within the subgroups were recalibrated separately to obtain item parameter estimates, and these latter estimates were combined into a second set, Set 2. It was found that the equating tables based on Set 1 were highly similar to those based on Set 2 item statistics. Although the IRT model theoretically requires one-dimensional tests, it appears to give satisfactory results with the LSAT. The equating tables appear to be adequate. Two appendixes contain LISREL computer program printouts for the factor solutions. (Contains 7 tables, 5 figures, and 26 references.) (SLD)

TM

The Effects of Dimensionality on True Score Conversion Tables for the Law School Admission Test

Gregory Camilli
Ming-mei Wang
Jaqueline Fesq

Law School Admission Council
Statistical Report 92-01
June 1992

LAW Services

A Publication of Law School Admission Council/Law School Admission Services

The Law School Admission Council is a nonprofit association of American and Canadian law schools. Law School Admission Services (Law Services) administers the Council's programs and provides services to the legal education community.

In this study, we examined the Law School Admission Test (LSAT) to see if the items on a form could be divided into different subgroups where items looked statistically similar within the subgroups, but statistically different between subgroups. When subgrouping can be detected, it is likely that the subgroups of items measure different abilities and therefore the test can be described as measuring multiple abilities or as "multidimensional" for short. In contrast, the term "unidimensional" is used to describe a test for which no subgroups exist—all the items look relatively similar from a statistical point of view. Such a test measures a single ability.

The LSAT is equated so that a test score obtained in the current year is comparable to scores obtained in previous years. Technically, a test model based on item response theory (IRT) is used to equate each new form of the LSAT to the base scale. This IRT model makes the basic assumption that the LSAT is unidimensional, as defined above. It is possible that a violation of this assumption could lead to unsatisfactory equating results; however, it must be recognized that 1) most tests are multidimensional to some degree, and 2) all practical test models for equating are unidimensional. Therefore, the two important issues with real tests concern the *degree* of multidimensionality of the test, and whether this has a practically significant effect on test equating.

To explore these issues, we conducted an analysis of multidimensionality for six forms of the LSAT using factor analysis. This statistical technique is commonly used to determine whether statistical subgroups of items exist, and which items correspond to which subgroups. We found two subgroups of items or "factors" for each of the six forms. The following pattern of results was remarkably consistent—the AR items corresponded to one factor, while the RC and LR items corresponded to the other. The main conclusion of the factor analysis component of this study was that the LSAT appears to measure two different reasoning abilities: inductive and deductive. Both RC and LR items appear to measure inductive reasoning, and AR items deductive reasoning. The item groupings identified are thus consistent with the content specifications of the LSAT. It is important to add that the analysis showed that these two reasoning abilities are highly correlated.

The technique of Dorans and Kingston (1985) was used in this study to examine the effect of dimensionality on equating. In brief, we began by calibrating (with IRT methods) all items on a form to obtain a set (say Set I) of estimated item parameters (*a*s, *b*s, and *c*s). Next, the test was divided into two homogeneous subgroups of items, each having been determined to represent a different ability (i.e., inductive and deductive reasoning). The items within these subgroups were then recalibrated separately to obtain item parameter estimates. These latter estimates were then combined into Set II. (All estimates were placed on the same scale.)

If the LSAT were strictly unidimensional, then the estimated item parameters in Set I would be very close to the corresponding estimates in Set II (only small differences would be obtained due largely to sampling errors). In other words, the same item statistics (*a*s, *b*s, and *c*s) would be obtained whether AR items were included with RC+LR items or not. Consequently, if the item statistics were the same, the equating tables based on parameter Sets I and II would be practically identical. On the other hand, if nonignorable multidimensionality exists, then the result of a single calibration of all LSAT items would differ noticeably from that of separate calibrations for the two subgroups of items. This could lead to different true score equating tables, depending on whether Set I or Set II item statistics were used.

In this study, we found that the equating tables based on Set I item statistics were highly similar to those based on Set II item statistics. We concluded, as did Dorans and Kingston (1985), that violations of unidimensionality may not have a substantial impact on equating. Although the IRT model theoretically requires unidimensional tests, it appears to give satisfactory results with the LSAT.

Contents

- The Effects of Dimensionality on True Score Conversion Tables for the Law School Admission Test 1
  - Section 1. Overview 1
  - Section 2. Description of the Data 4
  - Section 3. Pilot TESTFACT Runs 6
  - Section 4. Recent Work in Test Dimensionality 7
  - Section 5. Obtaining Inter-Item Correlation Matrices 8
  - Section 6. Design of Confirmatory Factor Solutions 11
  - Section 7. Results of Factor Analyses 12
  - Section 8. Results of Equating Analyses 16
  - Section 9. Future Research 22
  - Notes 22
  - References 23
- Appendix 1. Abridged LISREL Printout for the 3-factor Solution 25
- Appendix 2. Abridged LISREL Printout for the 11-factor Solution 27

Section 1. Overview

In this document, we first report the results of primary and secondary factor analyses of six forms of the Law School Admission Test (LSAT). Secondly, we report findings of how multidimensionality—as assessed by the factor analyses—may affect test equating. Finally, we consider potentially important areas for future research. Each of these topics is summarized briefly in this section, and is covered more fully in an ensuing section.

Factor Analyses

The objective of the factor analyses is to assess the dimensionality of LSAT items that are assembled into a test form. Though items for any particular form of the LSAT are administered in three sections—Analytical Reasoning (AR), Reading Comprehension (RC), and Logical Reasoning (LR)—two types of ability are included in the content specifications: inductive and deductive reasoning. It is highly probable that these item types and/or content domains are associated with statistical factors, and an analysis of dimensionality should reveal the extent of this association through a determination of 1) the number of underlying abilities or factors that adequately account for performance differences among examinees, and 2) the correlations among these abilities.

Model-based methods estimate factor structures directly from the data without resort to intermediate steps, i.e., correlations need not be computed. This is the strategy of full information (or IRT) factor analysis (Bock and Aitkin, 1981) which is available in the program TESTFACT (Wilson, Wood, and Gibbons, 1987). However, the full information approach was not used in this study because the AR and RC sections of the LSAT have a testlet structure (Wainer and Kiely, 1987), that is, a set of 4-8 items may pertain to a single passage. Higher correlations among items within testlets than among items between testlets resulted in extraneous (and tenuous) common factors in preliminary TESTFACT runs. To avoid this problem, a confirmatory approach was employed in this study to model item response dependencies.

Though a confirmatory approach was taken initially, an exploratory approach was employed subsequently to examine the common factor structure based upon correlations among testlets. It is thus useful to describe and distinguish these approaches. In this regard, Mulaik (1972) wrote

> In some situations where the researcher approaches a new domain, with practically no knowledge of what to expect, he will simply collect a representative sample of the variables to be measured in that domain and subject these variables to factor analysis. (Not very much in the way of a definitive theory about the variables should be expected to result from such use of factor analysis...) In other situations the researcher will have definite hypotheses as to the latent parameters present among the variables in a domain and will carefully select his variables so as to reveal the presence of the latent parameters as clearly as possible. In this case the researcher is involved to some extent in confirmatory research. (p. 362)

Mulaik also cited research (see pp. 363-366) indicating that in theoretically well developed areas "factor analysis does very poorly in recovering the already known theory." In this regard, there are two arguments for using confirmatory methods to analyze the dimensionality of LSAT items. First, the physical layout of the test items—which can be divided into analytical reasoning (AR), reading comprehension (RC), and logical reasoning (LR) sections—suggests there may be three factors, and that a search for dimensionality should begin (but not necessarily end) in this neighborhood. Second, the LSAT item domain is theoretically well developed in terms of its content, and this establishes useful prior expectations for the item structure. In particular, items appearing in the same passage should be more highly intercorrelated than cross-correlated with items of other passages, and passages should be correlated higher with other passages in the same content area than those in different content areas. A confirmatory approach can be most effectively "targeted" to these two prior expectations.

In the initial step of this study, correlation matrices were obtained for input to the general confirmatory factoring program LISREL 7 (Joreskog and Sorbom, 1989). However, with dichotomous data this presents a problem since we are interested in correlations of (continuous) latent variables underlying the propensity of examinees to answer test questions correctly. For this purpose, tetrachoric correlations were estimated (along with their error variances), rather than phi coefficients computed from observed responses, which have been shown to produce strong artifacts associated with item difficulty. Tetrachorics are less prone to this weakness; however, they tend to underestimate the latent correlations in the presence of guessing. This bias can be reduced by a simple adjustment described below. Although factoring tetrachorics is a piecemeal approach to assessing test structure (correlations are first estimated assuming both a multivariate distribution of latent variables and a parametric form for item response functions, and these estimates in turn are used to obtain estimates of model parameters), modern methods of factoring, as implemented by LISREL, appear to give results comparable to full information techniques (Bock, Gibbons, and Muraki, 1988).

In this study, tetrachorics were estimated from the full samples after an adjustment to the data for guessing. Upon obtaining the inter-item tetrachoric correlations and their variances, these matrices were input to LISREL to perform confirmatory factor analysis by the method of diagonally weighted least squares.[1] A number of different models were hypothesized:

1. A 3-factor solution was designed in which items from Analytical Reasoning (AR), Reading Comprehension (RC), and Logical Reasoning (LR) each loaded on separate factors.

2. An 11-factor solution was designed in which each testlet (defined as four or more test items subsumed under the same passage) of AR and RC items loaded on a separate factor, while LR items defined a single factor.

3. A 12-factor solution was designed with one general factor, one specific factor for each testlet.

4. A 13-factor solution was designed with two general factors and one specific factor for each testlet. The first general factor was constrained to have free loadings for AR items and loadings fixed at zero for RC and LR items. The second general factor was constrained by the opposite pattern of free and fixed loadings for the AR, RC, and LR items.

The 11-factor solution was found to fit significantly (and practically) better than the three 3-factor solution, while the 12- and 13-factor solutions fit marginally better than the 11-factor solution. We therefore focus on the results of the 11-factor solution in this report.

In the 11-factor model, correlations between the factors (i.e., the 10 testlets and the LR subtest) were estimated and then subjected to a second order factor analysis in order to discover whether the LSAT is unidimensional. Upon examining the results of the secondary analyses, it was concluded that a 2-factor second-order solution was appropriate. (Results from the IRT factor analyses, based on samples of 5,000 examinees, corroborated these results for one test form, but not another.) The following pattern of results was consistently obtained for six LSAT test forms: first, the AR testlets loaded highly on one factor, while the RC testlets loaded highly on the other; second, the LR subtest loaded on both factors, but more highly on the factor marked by the RC testlets; third, the estimated correlation among factors was about 0.7.

The main conclusion of the factor analysis component of this study was that the LSAT appears to measure two different reasoning abilities: inductive and deductive. One possible interpretation is that one factor (RC+LR) relates to the ability to understand relevancy and to make inferences, and the other (AR) relates to the ability to analyze a problem once it is discovered and defined. Another interpretation is that both RC and LR items appear to measure inductive reasoning, while AR items appear to measure deductive reasoning. The item groupings identified in the factor analysis are thus partially congruent with the physical structure of the test items, but more importantly, are wholly consistent with the content specifications of the LSAT. It is important to note that the analysis showed these two reasoning abilities are highly correlated. A correlation of 0.7 between the two second-order factors suggests that a single dominant second-order factor could account for 85% of the common factor variances.

Equating

The technique of Dorans and Kingston (1985) was used in this study to examine the effect of dimensionality on equating. In brief, this technique begins by calibrating (with IRT methods) an intact test to obtain a set (say Set I) of estimated item parameters (*a*s, *b*s, and *c*s). Next, the test is divided into two homogeneous subgroups of items, each having been determined to represent a different content domain (logically and statistically). The items within these subgroups are recalibrated separately to obtain sets of item parameter estimates (call these Sets IIa and IIb). Finally, all sets of parameter estimates are placed on a common metric, and Sets IIa and IIb are combined (call this Set II). By convention, this step is referred to as the "equating step" and the resulting ICC estimates are referred to as "equated ICCs."

If the test is strictly unidimensional, the estimated ICCs in Set I would be very close to the corresponding ICCs in Set II (small differences would be obtained largely due to sampling errors), and the true score equating tables based on parameter Sets I and II would be practically identical. On the other hand, if nonignorable multidimensionality exists, then the result of a single calibration of heterogeneous items (Set I) would differ noticeably from that of separate calibrations for the two groups of homogeneous items (Set II). In the presence of multidimensionality, the ICCs estimated in Set I would represent an ability that is qualitatively different from the two (correlated) latent abilities representing the ICCs in Sets IIa and IIb, respectively. Thus, the two sets of "equated" ICC estimates would likely differ, leading to different true score equating tables.

Based on the results of the factor analyses in the first phase of this study, the item parameters and abilities for a given form were estimated for

1. Heterogeneous (HT) sets of items. A heterogeneous set was represented by all AR, RC, and LR items on a given form. The term "heterogeneous" refers to the fact that we are combining items measuring different abilities. We also refer to the heterogeneous set as AR+RC+LR below.

2. Homogeneous (HM) subsets of items. There are two homogeneous subsets: the AR items, on one hand, and the RC and LR items on the other. We refer to these subsets as AR and RC+LR below, respectively.

The items in these sets were then calibrated with BILOG III (Mislevy and Bock, 1990) using default options, and were individually placed on the same scale using the characteristic curve method for scale transformation (commonly known as the *TBSE* procedure) described in Stocking and Lord (1983). In particular, the *a* and *b* estimates in Set I were linked to their operational form counterparts which had been previously placed on the LSAT base scale. This procedure was then repeated for Sets IIa and IIb, separately, and the results were then combined into Set II. In this way, all estimates were thus placed on the LSAT base scale, resulting in the comparability of Sets I and II.

The estimated *a*s, *b*s, and *c*s were highly similar for four LSAT forms. A correlational analysis showed that the correlations of *b*s from Sets I and II to be approximately 0.99 for the forms, while the correlations among *a*s ranged from 0.71 to 0.94 across forms. The correlations of *b* estimates *within* Sets AR and RC+LR to their AR+RC+LR counterparts were also high ($r \approx 0.99$). The correlations between *a*s within Set AR ranged from 0.62 to 0.92, and within Set RC+LR from 0.97 to 0.98. For the AR+RC+LR Set across the four forms, the root mean square difference for *a*s ranged from 0.07 to 0.14; for *b*s from 0.09 to 0.18; and for *c*s from 0.04 to 0.05.

The equivalence of the item parameter estimates was also examined within the context of the equating tables they produced. A true score conversion table was constructed for both the HT and HM calibrations separately using the June 1989 administration (0LSS2) as the base form. The HT and HM converted true scores were examined, and found to be highly similar. Throughout the true score range on form 0LSS2, the HT and HM calibrations resulted in true score conversion differences that ranged from -0.6 to +0.3 point. Thus, the effect of multidimensionality on true score conversions was found to be less than 1 point (or one question on form 0LSS2) for all four forms examined. We conclude, as did Dorans and Kingston (1985), that violations of unidimensionality may not have a substantial impact on equating. However, the effects on certain individuals may not be negligible. In future research we will attempt to identify these examinees.

Section 2. Description of the Data

The data come from six different administrations of the LSAT during the years 1989 and 1990. Each test is divided into three sections—Analytical Reasoning (AR), Reading Comprehension (RC), and Logical Reasoning (LR). The dates of administration, the total number of examinees at each administration, and the total number of scored items in each section of the test are given below in Table 1.

Table 1

Date	Form	Number of Examinees	Total Items Scored			Total Items
			AR	RC	LR	
June 1989	0LSS2	22088	29	34	33	96
Sept. 1989	0LSS1	43317	29	32	33	94
Dec. 1989	0LSS3	43796	29	34	31	94
Feb. 1990	0LSS5	29240	29	34	35	98
June 1990	1LSS7	25597	29	34	34	97
Oct. 1990	0LSS6	49644	29	34	35	98

The Analytical Reasoning section and the Reading Comprehension section both consist of five groups of "testlets" or passages. These are a series of four to eight items which all refer to the same reading passage or problem. The scoring for each item response is coded as 0 - incorrect, 1 - correct, 2 - omit, 3 - not reached, and 9 - not scored. Table 2 gives the average number correct, average number of omits, and average number of not reached items for each section of the test for each administration. It is evident from Table 2 that the average number of "omit" and "not reached" item responses across all three content areas is usually in the range of 1.0 to 1.5.

Before the correlations between items were calculated the scores for each person were adjusted for the omit and the not reached responses. A guessing factor of 0.20 was chosen because there are five possible choices for each item. If a randomly generated number between 0 and 1 was less than or equal to 0.20, the score of 2 (omit) or 3 (not reached) was recoded as a correct response. Otherwise, it was recoded as incorrect. This procedure is discussed more fully in the next section. (Statistics in Table 2 were calculated prior to random substitution.) Items that were not scored on final forms of the LSAT were omitted from the analyses.

Table 2

Descriptive statistics for omits, not reached, and number correct for six forms of the LSAT.

Variable	June 1989 (N = 22088) Mean	Std Dev	September 1989 (N = 43317) Mean	Std Dev	December 1989 (N = 43796) Mean	Std Dev
Analytical Reasoning						
Omits	.23	.90	.24	1.01	.23	.95
Not Reached	.40	1.75	.22	1.45	.26	1.76
Number Correct	18.14	5.52	17.98	5.65	18.98	5.59
Reading Comprehension						
Omits	.14	.77	.09	.52	.13	.68
Not Reached	.47	2.91	.29	2.28	.30	1.77
Number Correct	19.91	6.05	19.56	4.98	19.81	6.06
Logical Reasoning						
Omits	.09	.50	.08	.49	.12	.90
Not Reached	.30	2.64	.24	2.26	.14	.92
Number Correct	20.28	5.57	21.83	5.15	19.29	5.06

Variable	February 1990 (N = 29240) Mean	Std Dev	June 1990 (N = 25597) Mean	Std Dev	October 1990 (N = 49644) Mean	Std Dev
Analytical Reasoning						
Omits	.29	1.20	.20	.88	.19	.84
Not Reached	.36	2.21	.26	1.95	.14	.90
Number Correct	16.49	5.80	18.79	5.89	19.81	5.70
Reading Comprehension						
Omits	.16	.79	.11	.64	.08	.51
Not Reached	.36	1.77	.33	2.44	.24	2.15
Number Correct	19.93	6.82	21.50	6.86	23.10	5.82
Logical Reasoning						
Omits	.13	.62	.13	.65	.09	.51
Not Reached	.50	2.86	.26	1.30	.25	2.22
Number Correct	20.10	6.52	21.93	5.51	23.40	6.16

Section 3. Pilot TESTFACT Runs

As mentioned above, the LSAT has a testlet structure for the AR and RC subtests. Thus, there is good reason to believe a priori that statistical dependencies exist among items within testlets. For example, Thissen, Steinberg, and Mooney (1989) analyzed a 4-passage, 22-item test of reading comprehension in which the items appeared in clusters of 7, 4, 3, and 8. An item factor analysis with TESTFACT ($n = 3{,}866$) supported the existence of at least four factors, one of which was clearly a passage or testlet factor. In addition, there were two subpassage factors which were composed by the last few items in the first and fourth testlets. Thissen, Steinberg, and Mooney concluded

> Because there was one unambiguous passage factor, it appears that there are sometimes passage factors, which poses as much of a problem for the assumption of unidimensionality among reading comprehension items as if there were always passage factors. The subpassage factors also indicate that specific passage content influences the responses to some of the items following each passage. (pp. 249-250)

Two pilot runs with TESTFACT were done with the June (0LSS2) and September (0LSS1) 1989 LSAT data sets to explore the possibility of passage structures. For the June 1989 (sample $n = 5{,}000$) data, a 2-factor solution was found, indicated by a $X^2 = 12.24$ with 92 degrees of freedom ($p \approx 1.0$) for a potential third factor. The pattern of item loadings on the two factors (rotated to an oblique solution) was unambiguous: all AR items loaded on one factor, and all RC and LR items loaded on the other. However, one problem was noted with the solution. When residual correlations were inspected, it was clearly evident that large residuals were present within each cluster or testlet. This suggested the existence of passage clusters that were not detected by TESTFACT.

For the September 1989 data a 3-factor solution was found, indicated by a $X^2 = 775.76$ with 94 degrees of freedom ($p < 0.001$) for a third factor. (A 4-factor solution could not be run with the current PC version of TESTFACT.) The pattern of item loadings on the three factors (rotated to an oblique solution) was as follows:

1. Factor I was defined by a single RC testlet (items 15-21) concerning supply-side economics, other RC and LR items loaded moderately on this factor.
2. Factor II contained relatively high loadings for AR items with the exception of the first AR subtest. The RC items defining Factor I had moderately negative loadings on this factor.
3. Factor III was defined by a single AR testlet (items 1-6) concerning the composition of committees serving a university's board of trustees. Other AR items, as well as LR items, had low loadings on this factor. However, the RC testlet marking Factor I had loadings in the 0.12-0.20 range.

Confirmatory LISREL solutions were compared with the TESTFACT solution for the June and September 1989 data sets. The confirmatory solutions were more interpretable and stable across the forms. The LSAT has a passage structure that obstructs the identification of common factors by full information techniques as currently implemented. Consequently, it seemed preferable to use a confirmatory approach that takes into account prior information regarding the item structure of the test.

Section 4. Recent Work in Test Dimensionality

The assessment of test dimensionality is a topic of much interest in the current psychometric literature. McDonald (1981) suggested that under the assumption of unidimensionality, off-diagonal entries in the item variance-covariance matrix at different levels of examinee ability should be close to zero. Rosenbaum (1984) proved mathematically that multidimensionality implies that the inter-item covariance for any two items must be greater than zero for groups of examinees with identical scores on the remaining items. Stout (1987, 1990) broadened these ideas by examining the asymptotic behavior of inter-item covariances as the size of an item pool increased. (But note that an "item pool" is not synonymous with a "test form.") He coined the phrase *essentially unidimensional* to refer to a test with one major factor and one or more minor factors. Both Rosenbaum and Stout suggested statistical procedures for dimensionality testing.

Reckase (1979) concretely demonstrated that 2- and 3-parameter IRT models can consistently retrieve the major dimension of an essentially unidimensional test. However, he also showed that for a test with equally potent dimensions, IRT procedures provided inconsistent estimates of ability. Most useful tests are not unidimensional, yet may be essentially so. Hence, unidimensional test models and procedures may give satisfactory results. Hambleton (1989, p. 150) wrote that "What is required for the assumption of unidimensionality to be met to a satisfactory extent by a set of test data is a *dominant* component or factor."

Dominance is presumably measured by the "fit" of a unifactor model relative to a multifactor model; however, this analysis may oversimplify the problem of assessing dimensionality. Other aspects of dimensionality must be considered because models that are logically quite different can often be constructed to fit data structures equally well. Thus, indices of statistical fit should not be the sole arbiters of dimensionality. Dimensionality is not a property of a test per se—it is context dependent. Two short examples may serve to illustrate problems with a purely "objective" approach to dimensionality assessment. Zwick (1987) noted that reliance on full information fit statistics tended to lead to overfactoring, and therefore concluded that "the size of factors and the patterns of loadings should also be considered in determining the number of factors." For Stout's (1987) test of dimensionality, a short homogeneous set of assessment items must be chosen, either by expert opinion or the strategic use of factor analysis (Nandakumar, 1991). Note that both of these procedures require some substantive knowledge on the part of the researcher in interpreting patterns of loadings or selecting a set of homogeneous items.

We think that "test dimensionality" is a type of validation argument concerning a test related to content validation. It is well-recognized that content validity is not a property of test items themselves, but rather a function of the interaction of examinees with test items *and* a function of the test's use. In this regard Messick (1989, p. 41) wrote

> Strictly speaking, even from a content viewpoint, it would be more apropos to conceptualize content validity as residing not in the test, but in the judgment of experts about domain relevance and representativeness.

Because the "nature and dimensionality of the inter-item structure should reflect the nature and dimensionality of the construct domain," (Messick, 1989, p. 44) it is likely that judgments regarding the content affect those regarding the dimensionality of a test. Therefore, test content is not subordinate to test dimensionality; rather, an argument must be made for a particular conclusion regarding dimensionality, and this argument should incorporate both judgments about test content and evidence from statistical analyses. In short, arguments concerning dimensionality need to be validated in the context of a test's use.

In the analyses below, we incorporate judgments concerning the content domain of the LSAT into a statistical analysis of item structure. We conceive of this approach as an analysis of *functional dimensionality* rather than one of *statistical dimensionality*. This approach is motivated by the need to manage the item dimensionality as opposed to the unrealistic goal of creating a unidimensional test. The practical consequences—to both the developer and examinee—are assessed in this study of treating a multifactor test as unidimensional for the purpose of equating.

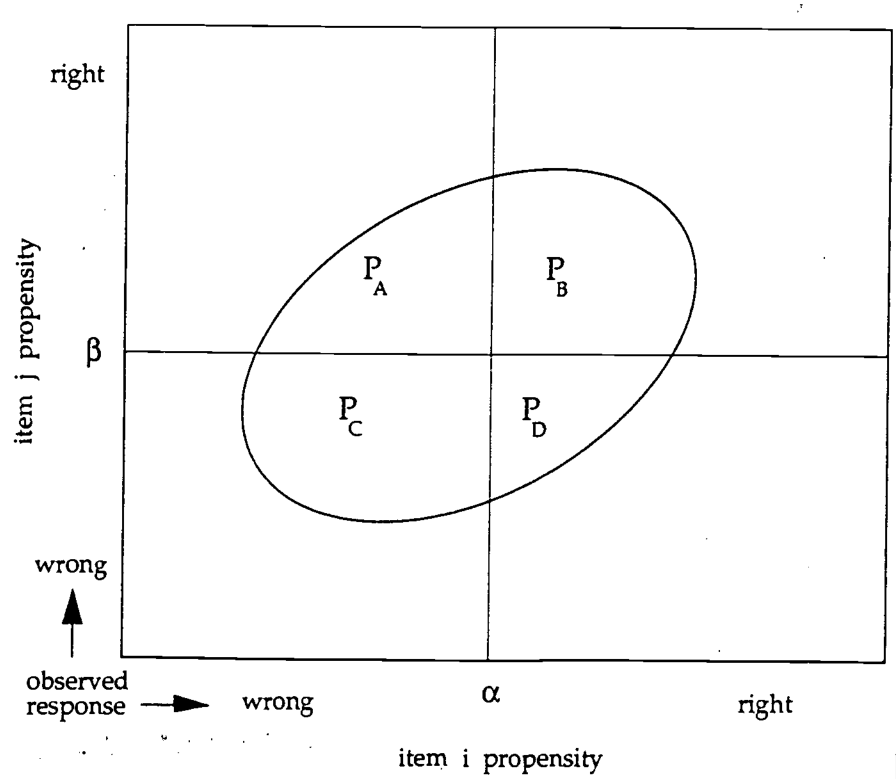
Section 5. Obtaining Inter-Item Correlation Matrices

The factoring methods used in the next section are based on inter-item tetrachoric correlations. Though tetrachoric correlations may yield spurious factors (Carroll, 1983; Hulin, Drasgow, and Parsons, 1983), there is one approach that seems to yield acceptable results. Christoffersson (1975) and Muthen (1978) developed a generalized least squares (GLS) approach based on the work of Browne (1984). Bock, Gibbons, and Muraki (1988) compared a full information solution with the corresponding GLS solutions. The results were highly similar leading to the conclusion that the essential correctness of both methods was supported. However, the GLS solution requires a large amount of computer memory and is practically implemented for only about 25 items. A compromise approach, diagonally weighted least squares, has been developed and is economically implemented. Below, we provide a rationale for the use of tetrachoric correlations. The method used to factor analyze matrices of these coefficients will be described in Section 6.

In modern measurement theory, item responses depend linearly on the true ability (or abilities) of the examinee, an item "threshold," and an error of measurement. When the propensity of an examinee to get an item correct exceeds this threshold, a correct response or "1" is observed. If it falls below this threshold, an incorrect response or "0" is observed. An examinee's true ability and propensity are unobserved variables, only the dichotomous response is observed. This suggests two general strategies for estimating inter-item correlations: to use observed responses in the standard correlation formula; or to assume a model by which discrete responses are generated from continuous responses, and to estimate a correlation within the framework of this model. For example, if it is assumed that the (unobserved) propensities to answer any pair of items correctly are bivariate normal, then the tetrachoric correlation r_t is defined as the correlation parameter in the bivariate normal distribution function.

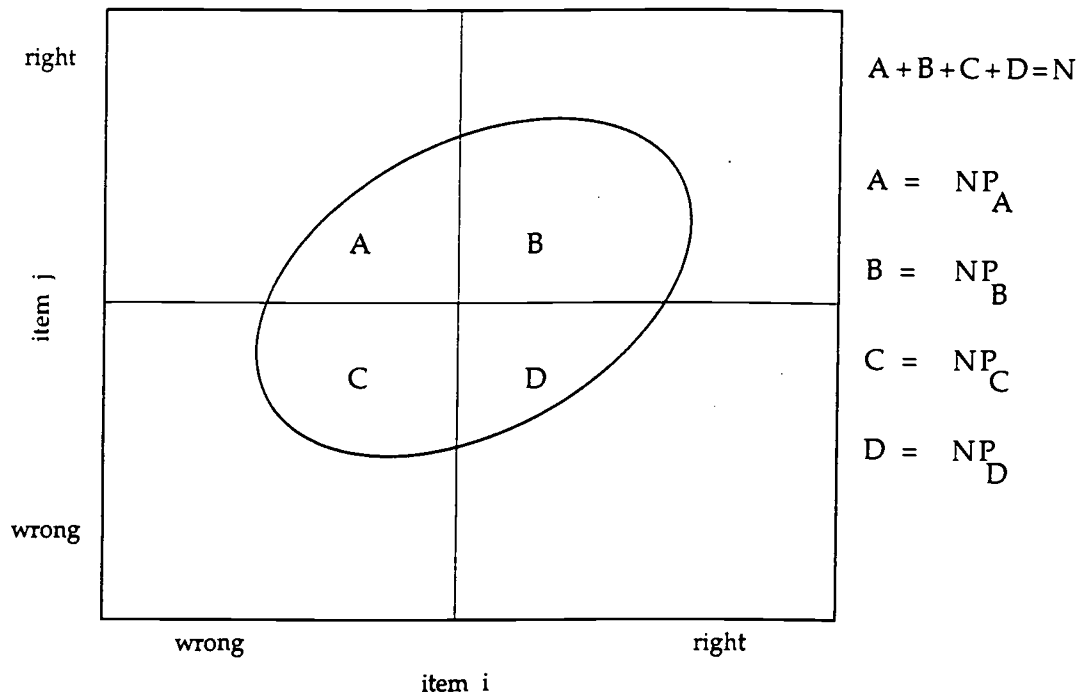
The latter approach is diagramed in Figure 1a, in which a bivariate normal distribution of propensities is dichotomized at points α and β. This process leads to the observed frequencies in the 2x2 contingency table in Figure 1b for a test population with N examinees.

Figure 1a
Bivariate normal model for generating observed dichotomous item responses.



BEST COPY AVAILABLE

Figure 1b
Crosstabulated frequencies of observed correct and incorrect responses from the bivariate normal model.



To estimate r_t, we work backward from the observed data to the parameters of the bivariate function that "generates" the data. The method of maximum likelihood (ML) estimation selects r_t as the value that makes the observed (tabled) data "most likely." For example, it would be extremely improbable that inter-item correlations of 0.0 or 1.0 would generate the responses that are observed from most real test items because it is common to see some examinees miss one item of any pair while answering the other correctly. It is more likely that such observed item responses conform to some intermediate point in the interval (0.0, 1.0). In this study, the results by Tallis (1962) were used to write an ML algorithm for estimating r_t for all pairs of items. The conditional approach was chosen in which the marginal means of the bivariate distribution of propensities (α and β) were fixed. Simultaneous ML estimation of all parameters was attempted, but severe convergence problems were encountered in a number of instances. Tallis (1962) noted other drawbacks of simultaneous estimation.

Entries in the 2x2 table were adjusted for guessing prior to ML estimation. Briefly, this entailed shifting a proportion of responses from correct to incorrect to compensate for guessing. Carroll (1945, 1983) and Bock, Gibbons, and Muraki (1988) presented a more complete treatment of this adjustment. A common guessing parameter of 0.1 was chosen for all items. This is the approximate value of the median *c* estimate on LOGIST calibrations of the LSAT (Alison Sneickus, personal communication). The error variances of the r_ts were also estimated with the ML algorithm, and were saved for further use (as described below). Because some correlations are estimated much more precisely than others as indicated by smaller variances, it is desirable to give such estimates more weight in ensuing analyses. All correlation matrices were obtained from the full population of test takers, and not small subsamples (sample sizes ranged from 25,000 to 45,000).

In a number of instances examinees failed to attempt or did not reach an item. Such item responses are scored as incorrect which results in mis-estimation of the r_t (Carroll, 1945, 1983). To compensate for this source of error, we substituted plausible responses for omitted and not-reached items, and then employed Carroll's correction for guessing uniformly for all examinees. Specifically, we assumed that if these examinees followed the instructions given in the LSAT information booklet, they would quickly make guesses for omitted or unreached items at the end of the test period. This guessing process would result in a probability of 0.2 (with a five option item) of answering correctly, while a "normal" guessing process would still result in a probability of 0.10. Thus, when omitted or unreached responses were encountered, item responses were randomly generated according to the following model. A uniform random deviate U was drawn on the interval [0,1]. If $U \leq 0.2$, then a "1" was recorded; otherwise, a "0" was recorded.

Section 6. Design of Confirmatory Factor Solutions

Item factor analysis typically is encountered in two forms, exploratory and confirmatory. The confirmatory use is characterized by imposing to (a limited degree) an hypothesized structure on the item data (in the form of the inter-item correlation matrix) and measuring the "fit" or how well the structure accounts for the data. Since the "true" structure is rarely, if ever known, the basic strategy of the confirmatory approach is to compare different or rival hypotheses about the structure. If one structure "fits" better than another, in terms of both practical and statistical significance, than it is preferred. Clearly, there is an intended structure to the LSAT which is based both on content specification and item type. The LSAT is divided into three main sections labeled Analytical Reasoning (AR), Reading Comprehension (RC), and Logical Reasoning (LR). Items in these sections may be self-contained or may appear in clusters. Item clusters, or testlets, are items that refer to the same extended stem or passage. All AR and RC items are clustered in groups of 4-8, while LR items most frequently stand alone (or infrequently in couplets).

Two preliminary confirmatory solutions were designed: first, a 3-factor solution in which items from Analytical Reasoning (AR), Reading Comprehension (RC), and Logical Reasoning (LR) each loaded on separate factors; and second, an 11-factor solution in which each testlet (defined as four or more test items subsumed under the same passage) of AR and RC items loaded on separate factors, while LR items defined a single factor. Within both designs, correlations among factors were estimated as *free parameters*. Upon estimation, the testlet (or factors representing testlets) correlation matrix was obtained for both designs. The procedures for the 3- and 11-factor solutions were thus confirmatory except for the fact that a correlational structure among the factors was not directly imposed.

For the 11-factor solution, we expected that passages should be correlated higher with other passages in the same content area than those in different content areas. To reveal this potential phenomenon, we designed a secondary confirmatory analysis for the 11-factor correlation matrix. In this design, we assumed two factors. The first was identified by fixing the loadings of one AR testlet to "1" on the first factor and "0" on the second. Likewise, the loadings for one RC testlet was fixed as "0" on the first factor and "1" on the second. Aside from these constraints, all other coefficients were treated as free parameters including the correlation of the two factors. Finally, the standardized solution for this design was obtained with LISREL.

This two step process was opted instead of a completely exploratory solution (in which no structure is imposed except for the number of factors) for two reasons. First, exploratory approaches rarely identify correct models in the presence of highly correlated variables. Such models efficiently provide the "best fit" to the data, but in a predictive rather than substantive sense. In fact a number of models may fit the data reasonably well so that other criteria such as interpretability, parsimony, and cross-validation should also be considered. Second, a confirmatory approach based on the content specification of the LSAT more closely relates the statistical structure of the items to the item development process. For example, examining how a testlet functions within a multidimensional test is more relevant to item developers than examining a single item from a testlet. This is because items from clusters are not developed independently, nor are they independent in a statistical sense. Factoring methods that recognize this non-independence are likely to have greater practical significance as well as providing a good fit to the data.

For estimation, a diagonally weighted least squares (DWLS) algorithm was used (Joreskog and Sorbom, 1989). This method iteratively approximates model coefficients (i.e., factor loadings) by obtaining an initial solution and then using this as a stepping-stone to locate a "better" solution. Here a "better" solution means one in which the tetrachoric correlation matrix can be reproduced more accurately from estimated factor coefficients. (Because the factor coefficients are considered to "generate" the correlational structure.) During estimation, the sum of squared discrepancies (say **S**) is successively reduced between inter-item correlations (say $\hat{r}_{gh}$) predicted by a particular factor structure and the observed tetrachorics (say r_{gh}). Technically, a new solution is chosen at step i+1 that shows better fit than the solution at step i, thus the strategy is to minimize **S**. Iterations are ended when decreases in **S** become negligible.

The method of DWLS defines **S** as

$$S = \sum_{g \neq h} \left(\frac{1}{w_{gh}}\right)\left(\hat{r}_{gh} - r_{gh}\right)^2$$

where w_{gh} is the conditional asymptotic variance of the ML estimate r_{gh}. In this manner, estimates of r_{gh} with more sampling error contribute less to the solution. Both the rs and their weights were obtained with the full samples of test takers (see Table 1 for sample sizes) and then input to the DWLS algorithm implemented in LISREL 7. This method is more efficient than an unweighted solution (Joreskog and Sorbom, 1989). (The method of fully, or generally weighted least squares, which does yield asymptotically efficient estimates, is not practically implemented with more than 20-30 items.)

Section 7. Results of Factor Analyses

First-order Analyses

Six forms of the LSAT were factored using confirmatory methods, and model coefficients were obtained. These included factor loadings and inter-factor correlations. The solutions were highly similar for each form; therefore, the LISREL analyses are given for the September 1989 administration only (0LSS1). For the 3-factor solution of this data set (Appendix 1), the adjusted goodness of fit index (AGI) and root mean square residual (RMSR) were 0.963 and 0.040, respectively. For the 11-factor solution (Appendix 2), the AGI and RMSR were 0.986 and 0.027. For the purpose of interpretation, AGI is an index that varies between 0 and 1, with 1 being "perfect" fit, while the RMSR index should be close to 0 with models that fit the data closely. The X^2 values for these models were not meaningful given the large sample sizes—all X^2 values obtained were on the order of 10^6. The AGI and RMSR were used to *compare* the fits of the two models. For example, the 11-factor solution had an AGI about 2% higher (0.986 v 0.963), and an RMSR about 50% lower (0.027 v 0.040) than the 3-factor model. This finding was consistent for each of the six forms examined. Thus, we concluded that the 11-factor solution gave a meaningfully better fit.

For the 3-factor solution it was observed that the correlations across the six LSAT forms between the RC and LR factors averaged 0.9, while the AR/RC and AR/LR correlations averaged 0.66 and 0.75, respectively. Thus, it appeared that a 2-factor solution might be appropriate with RC and LR defining one factor, due to their high intercorrelation, and AR defining another. Two additional confirmatory models were examined for determining the number of factors. For 2 LSAT administrations (0LSS2 and 0LSS3) a 12-factor solution was obtained with one general factor and 11 unique factors, and a 13-factor solution was obtained with two general factors and 11 unique factors. Both of these models fit marginally better than the 11-factor solution; however, the 13-factor model fit better than the 12-factor model (in terms of the AGI and RMSR) suggesting the presence of two factors.

The 11-factor design, or testlet design, probably acquired its relative power by modeling the interdependence of items within a testlet. Such items were not locally independent in the 3-factor solution due to the same characteristic of the overarching passage: inter-item correlations within a passage were significantly underestimated. For example, item responses pertaining to a passage dealing with supply-side economics showed very high residual correlations for the 3-factor model. Nonsubstantive characteristics of some passages could also have lead to local dependence. In the testlet design, each testlet defined an ability that combined both common (say reading comprehension) and unique (say knowledge of economics) elements. The unique elements may have affected item responses even though they were not strictly required for answering correctly; for example, substantive familiarity could have increased processing efficiency.

Second-order Analyses for the 11-factor Model

As mentioned above, we expected that passages should be correlated higher with other passages in the same content area than those in different content areas. To reveal this potential structure without imposing it, we designed a secondary analysis (see Section 6) with only mild restrictions on the factor structure. The standardized solutions for this design were obtained for 3 LSAT administrations: June (0LSS2) and December (0LSS3) 1989, and October (0LSS6) 1990. These solutions and the 11-factor correlation matrices appear in Tables 3-5.

The findings were highly similar across these three analyses. A highly consistent structure emerged in which Factor 1 was marked by the AR testlets and to a slight degree by LR, and Factor 2 was marked by the RC testlets and to a lesser degree by the LR subscale. These factors, which were labeled RC/LR and AR, were moderately correlated (0.56 - 0.74). Thus, the two major abilities influencing test performance are not statistically independent. The RC and LR abilities may be distinct, but correlated to such a high degree that they were practically indistinguishable. However, because both RC and LR items are also *intended* to measure inductive reasoning, we conclude that one statistical factor subsumes both item types. Finally, it is noted that the secondary factor results were corroborated by purely exploratory factor analyses of the inter-factor correlation matrices.

Table 3

Factor inter-correlation matrix and secondary factor analysis results for the June 1989 administration (0LSS2).

Correlation Matrix

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
Var 2	0.477									
Var 3	0.473	0.577								
Var 4	0.389	0.503	0.634							
Var 5	0.442	0.472	0.588	0.653						
Var 6	0.453	0.446	0.490	0.421	0.479					
Var 7	0.420	0.473	0.490	0.432	0.475	0.818				
Var 8	0.328	0.310	0.336	0.319	0.338	0.544	0.545			
Var 9	0.385	0.404	0.441	0.433	0.449	0.751	0.736	0.519		
Var 10	0.437	0.419	0.487	0.503	0.597	0.768	0.731	0.483	0.764	
Var 11	0.558	0.559	0.611	0.604	0.637	0.838	0.838	0.575	0.789	0.836

Standardized Factor Solution

	Factor 1	Factor 2
Var 1	0.432	0.206
Var 2	0.594	0.090
Var 3	0.777	0.000
Var 4	0.889	-0.113
Var 5	0.757	0.035
Var 6	-0.083	0.959
Var 7	-0.052	0.925
Var 8	0.000	0.603
Var 9	-0.056	0.881
Var 10	0.116	0.778
Var 11	0.249	0.771

Factor Correlation r = .713

Root Mean Square Residual .025

Table 4

Factor inter-correlation matrix and secondary factor analysis results for the September 1989 administration (0LSS1).

Correlation Matrix

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
Var 2	0.522									
Var 3	0.521	0.683								
Var 4	0.493	0.699	0.672							
Var 5	0.414	0.593	0.625	0.641						
Var 6	0.510	0.574	0.535	0.599	0.477					
Var 7	0.393	0.483	0.404	0.496	0.355	0.790				
Var 8	0.383	0.483	0.400	0.537	0.373	0.779	0.798			
Var 9	0.480	0.590	0.581	0.611	0.531	0.804	0.729	0.739		
Var 10	0.372	0.489	0.484	0.540	0.502	0.688	0.659	0.642	0.789	
Var 11	0.556	0.653	0.599	0.680	0.551	0.873	0.791	0.812	0.864	0.741

Standardized Factor Solution

	Factor 1	Factor 2
Var 1	0.500	0.187
Var 2	0.741	0.123
Var 3	0.834	0.000
Var 4	0.720	0.170
Var 5	0.746	0.013
Var 6	0.191	0.789
Var 7	-0.018	0.883
Var 8	0.000	0.878
Var 9	0.291	0.700
Var 10	0.234	0.627
Var 11	0.307	0.752

Factor Correlation r = .558

Root Mean Square Residual .021

Table 5

Factor inter-correlation matrix and secondary factor analysis results for the October 1990 administration (0LSS6).

Correlation Matrix

	Var 1	Var 2	Var 3	Var 4	Var 5	Var 6	Var 7	Var 8	Var 9	Var 10
Var 2	0.592									
Var 3	0.633	0.577								
Var 4	0.645	0.645	0.665							
Var 5	0.550	0.563	0.533	0.701						
Var 6	0.538	0.494	0.530	0.551	0.485					
Var 7	0.555	0.511	0.567	0.556	0.489	0.905				
Var 8	0.518	0.449	0.511	0.504	0.426	0.827	0.853			
Var 9	0.494	0.452	0.526	0.502	0.436	0.781	0.814	0.769		
Var 10	0.464	0.418	0.483	0.475	0.440	0.700	0.718	0.711	0.709	
Var 11	0.645	0.573	0.665	0.658	0.578	0.835	0.896	0.836	0.818	0.781

Standardized Factor Solution

	Factor 1	Factor 2
Var 1	0.772	-0.008
Var 2	0.811	-0.098
Var 3	0.776	0.000
Var 4	1.008	-0.192
Var 5	0.872	-0.159
Var 6	0.048	0.884
Var 7	0.044	0.927
Var 8	0.000	0.896
Var 9	0.060	0.813
Var 10	0.120	0.690
Var 11	0.324	0.686

Factor Correlation r = .745

Root Mean Square Residual .016

Section 8. Results of Equating Analyses

For 4 LSAT forms (0LSS2, 0LSS3, 0LSS5, and 0LSS6), items were calibrated with BILOG using standard options. The estimated *as*, *bs*, and *cs* were obtained for heterogeneous (HT) and homogeneous (HM) groups of items, and placed on the same scale with the TBSE procedure. The item parameter estimates for the HM and HT calibrations were then compared by examining correlations and root mean square differences (RMSD). The results were similar across forms, and are typified by those for 0LSS2 and 0LSS3 which are given in Tables 6 and 7.

Overall, several observations can be made. First, the correlations among the HM and HT estimated *bs*, whether in AR or RC+LR sets of items, were quite high with a modal value of about 0.99. Second, the correlations among the *a* and *c* parameter estimates were high for RC+LR sets, but lower for AR sets. Third, the differences between the HM and HT parameter estimates were relatively low for the RC+LR subset. For example, the 0LSS2 root mean square differences (RMSD) for RC+LR were 0.04, 0.08 and 0.02 for the *a*, *b*, and *c* differences, respectively, while these differences for the AR set were 0.10, 0.30, and 0.07. In general, the RMSDs were about twice as high for the AR sets as the RC+LR sets.

Table 6

Correlations and root mean square differences for HM and HT item parameters from form 0LSS2.

	A2	B2	C2
All Items			
Correlations			
A1	.92**	.04	.09
B1	.10	.99**	.18
C1	.27**	.10	.86**
RMSD	.07	.18	.04
AR Items			
Correlations			
A1	.73**	.40*	.18
B1	.60**	.98**	.30
C1	.46**	.15	.74**
RMSD	.10	.30	.07
RC + LR Items			
Correlations			
A1	.97**	-.07	-.09
B1	-.07	.99+**	.14
C1	-.03	.10	.96**
RMSD	.04	.08	.02

* p < .05
**p < .01

Table 7

Correlations and root mean square differences for HM and HT item parameters from form 0LSS3.

All Items

	A2	B2	C2
Correlations			
A1	.94**	.27**	.06
B1	.26*	.99**	.15
C1	.04	.12	.86**
RMSD	.09	.15	.05

AR Items

	A2	B2	C2
Correlations			
A1	.91**	.42*	.03
B1	.53**	.98**	.33
C1	.14	.36	.77**
RMSD	.10	.22	.07

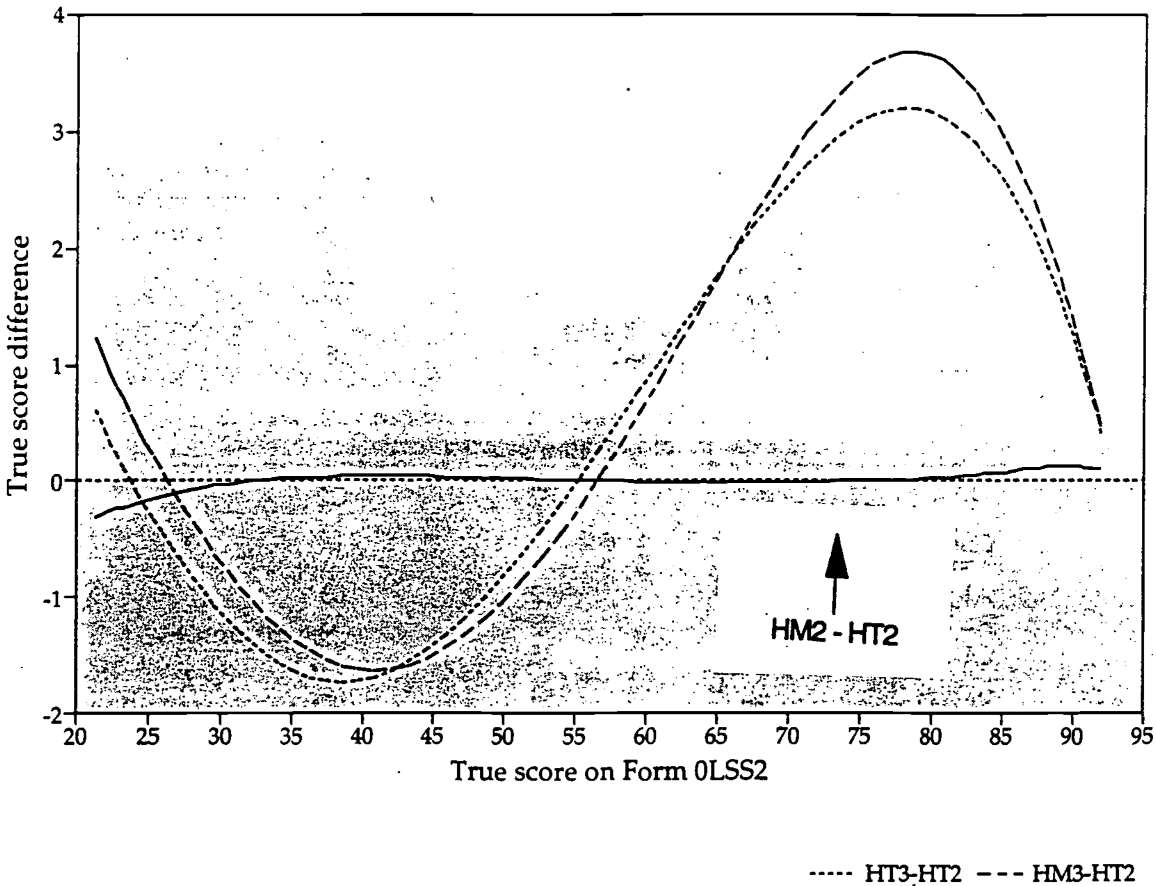
RC + LR Items

	A2	B2	C2
Correlations			
A1	.97**	.24*	.03
B1	.26*	.99+**	.09
C1	.03	.03	.97**
RMSD	.06	.08	.02

* $p < .05$
** $p < .01$

The equivalence of the true score equating tables was also examined for the HM and HT item parameter estimates. A conversion table was constructed for both the HT and HM calibrations separately for converting true scores for four LSAT administrations to true scores on a base form (0LSS2). The base form scale was defined by the simultaneous calibration of all items (i.e., the HT set), and conversions were plotted as the *difference* between the HT or HM calibrations for each form and the HT calibration for the base form (labeled HT2). In Figure 2, the results are presented for converting 0LSS3 scores to the 0LSS2 scale.

Figure 2.
Score conversions for 0LSS3.



The HM conversion function to base is given by HM3-HT2, and the HT conversion to base is given by HT3-HT2. As can be seen, the conversion functions for the heterogeneous and homogeneous calibrations, though nonlinear, are highly similar: throughout the true score range on form 0LSS2, the two calibrations appear to result in true scores that differ at most by about one half point. Also in Figure 2, it can be seen that true scores for the HM and HT calibrations of the base form (HM2 and HT2) are also similar having a difference (HM2-HT2) close to zero throughout the true score range.

Similar conversion functions are plotted in Figures 3 and 4 for forms 0LSS5 and 0LSS6. These HM and HT conversions also resulted in small differences.[2] In Figure 5, only the HM-HT differences are plotted for each form. These differences ranged from -0.6 to +0.3 point. The largest differences were observed at the low and high regions of the base form scale. Still, the effect of multidimensionality on true score conversions is less than 1 point, or one question on form 0LSS2. Therefore, violations of unidimensionality as evidenced by the factor analyses do not appear to have a substantial impact on true score conversion tables.

Figure 3.
Score conversions for 0LSS5.

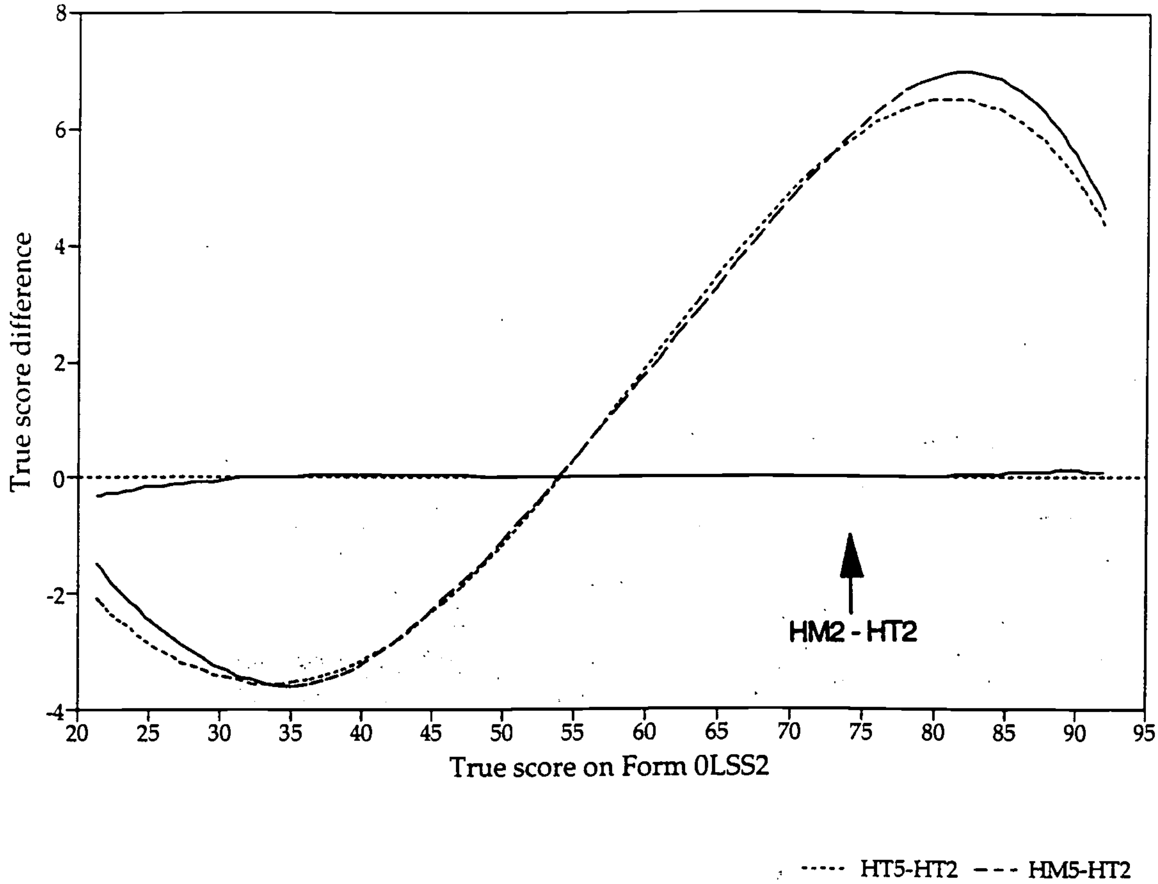
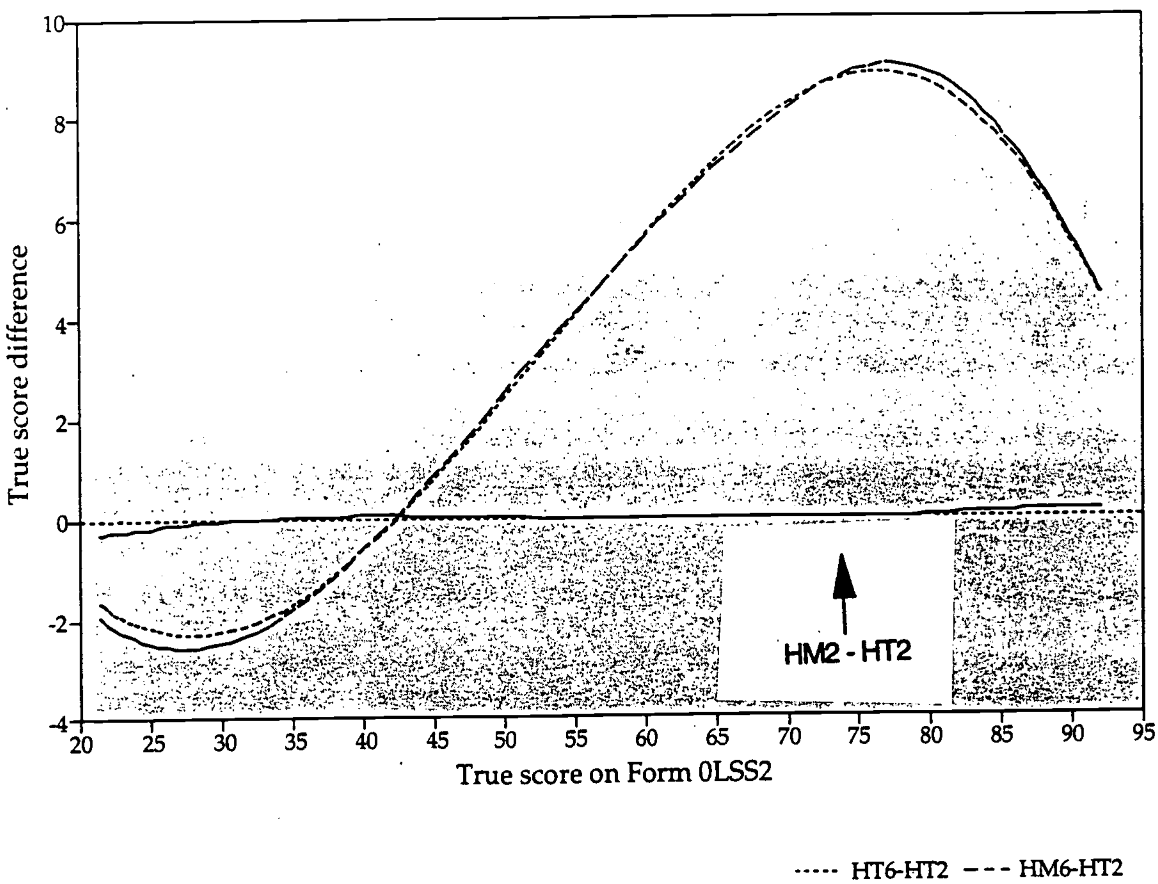
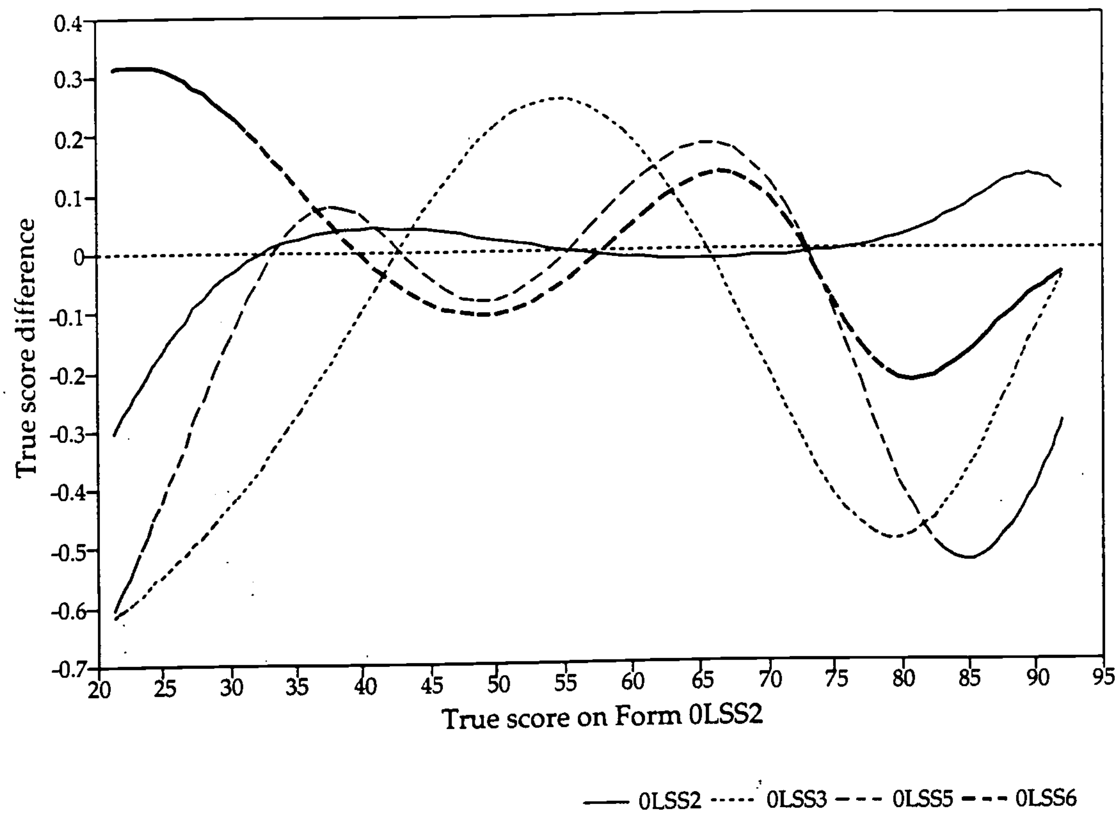


Figure 4.
Score conversions for 0LSS6.



BEST COPY AVAILABLE

Figure 5.
Differences between true scores for homogeneous and heterogeneous item calibrations for four test forms.



BEST COPY AVAILABLE

Section 9. Future Research

The effect of multidimensionality on the true score conversion tables appears to be minimal. However, inherent in the method of Dorans and Kingston (1985) is the assumption that the quantity being equated is a composite of underlying abilities. This is because the LSAT raw score, being the sum of right/wrong item responses, forces a unit weighting of items that may represent different underlying factors. An alternative scoring method based on a 3P IRT model would also extract a composite of the underlying multiple abilities, though an item's contribution to ability estimation would be weighted. We have found in this study that true score conversions for the composite score are not greatly affected by the presence of two substantially correlated factors.

The effect of multidimensionality could be substantially different for methods of equating that do not use a composite ability. For example, suppose true score conversion tables were created for AR and RC+LR separately. The "total score" on the base scale could then be obtained as the sum of the two converted true scores. Now the question could be asked "What is the difference between the heterogeneous test conversion and the sum of the two homogeneous test conversions?" An analysis of this phenomenon is more complex than the analysis in this report for two reasons. First, there is not a single conversion function; rather, there is a converted true score on the base form for each pair of true scores on the new form. Thus, the equating function is three dimensional requiring three dimensional versions of Figures 2-5. Second, if two subtests were defined for equating, each would be shorter than the full test resulting in diminished reliability. Also note that the AR subtests (with 29 items) would have substantially less reliability than the RC+LR subtest (with 63-69 items).

We have implied above that estimates of ability on the AR and RC+LR subtests may be substantially different for some examinees. This possibility is suggested by the moderate correlation ($r \approx 0.7$) between the two abilities. It seems important for two reasons to identify the examinees for whom this discrepancy is large, and to determine whether the discrepancy is a function of demographic or other background variables. First, systematic discrepancies in ability estimates for certain groups creates the potential for differential item functioning. Second, if both abilities are important for the prediction of academic performance, then some loss of predictive power might be associated with treating the test as unidimensional.

Notes

1. Often matrices of estimated tetrachorics are nonpositive definite which is problematic for statistical methods that require inversion of this matrix. In the present study, all tetrachoric matrices were positive definite, though the method of diagonally weighted least squares implemented in LISREL 7 does not require this condition.
2. The amplitude of the S-shaped curves, i.e., the vertical distance of a single curve to the zero point on the Y-axis, is affected by the number of items on the tests being equated. For this reason, it is important to focus on the *difference* between the two equating curves for the purpose of assessing the effects of multidimensionality.

References

Bock, R.D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: An application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R.D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. Applied Psychological Measurement, 12, 261-280.

Browne, M.W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. British Journal of Mathematical and Statistical Psychology, 37, 62-83.

Carroll, J.B. (1945). The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 10, 1-19.

Carroll, J.B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (Eds.) Principals of modern psychological measurement (pp. 257-282). Hillsdale, NJ: Lawrence Erlbaum.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Dorans, N.J., & Kingston, N.M. (1985). The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE verbal scale. Journal of Educational Measurement, 22, 249-262.

Hambleton, R.K. (1989). Principles and selected applications of item response theory. pp. 147-200. In R.L. Linn (Ed.), Educational Measurement (third edition), New York: Macmillan.

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

Joreskog, K.G., & Sorbom, D. (1989, 2nd ed.). LISREL 7: A Guide to the Program and Applications. Chicago, IL: SPSS, Inc.

Messick, S. (1989). Validity. pp. 13-103. In R.L. Linn (Ed.), Educational Measurement (third edition), New York: Macmillan.

Mislevy, R.J. & Bock, R.D. (1990). BILOG 3.04: Item Analysis and Test Scoring with Binary Logistic Models. Mooresville, IN: Scientific Software, Inc.

Mulaik, S. A. (1972). The foundations of factor analysis. New York: McGraw-Hill.

Muthen, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

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

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

Reckase, M.D. (1979). Unifactor latent trait models applied to multifactor tests: results and implications. Journal of Educational Statistics, 4, 207-230.

Rosenbaum, P.R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. Applied Psychological Measurement, 7, 201-210.

Stout, W. (1987). A nonparametric approach for assessing latent trait dimensionality. Psychometrika, 52, 589-618.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. Psychometrika, 55, 293-325.

Tallis, G.M. (1962). The maximum likelihood estimation of correlation from contingency tables, Biometrics, 18, 342-353.

Thissen, D., Steinberg, L., & Mooney, J. (1989). Trace lines for testlets: a use of multiple-categorical-response models. Journal of Educational Measurement, 26, 247-260.

Wainer, H., & Kiely, G.L. (1987). Item clusters and computerized adaptive testing: a case for testlets. Journal of Educational Measurement, 24, 185-201.

Wilson, D., Wood, R.L. & Gibbons, R. (1987). TESTFACT: Test scoring and item factor analysis. Mooresville, IN: Scientific Software.

Zwick, R. (1987). Assessing the dimensionality of the NAEP reading data. Journal of Educational Measurement, 24, 293-308.

Appendix 1.
Abridged LISREL Printout for the 3-factor Solution

LISREL 7:
Estimation of Linear Structural Equation Systems
Program Version 7.16

Distributed by:
Scientific Software, Inc.
1369 Neitzel Road
Mooresville, Indiana 46158
(317) 831-6336

MVS - L I S R E L 7.16
By
Karl G. Joreskog and Dag Sorbom

The following LISREL control lines have been read:

```
DA NI=94 NO=43317 MA=PM
PM UNIT=18
DM UNIT=17
MO NX=94 NK=3 LX=FR TD=DI,FR PH=SY
FI PH(1,1) PH(2,2) PH(3,3)
VA 1.0 PH(1,1) PH(2,2) PH(3,3)
PA LX
29*(1 0 0 ) 32*(0 1 0) 33*(0 0 1)
OU DWLS TM=1500
September 1989. Three factor solution.
```

Number of Input Variables	94
Number of Y-Variables	0
Number of X-Variables	94
Number of ETA-Variables	0
Number of KSI-Variables	3
Number of Observations	43317

Warning: Chi-square, standard errors, t-values and standardized residuals are calculated under the assumption of multi-variate normality.

LISREL Estimates (Diagonally Weighted Least Squares)
Lambda X

	KSI 1	KSI 2	KSI 3
Var 1	0.627	0.000	0.000
Var 2	0.607	0.000	0.000
Var 3	0.569	0.000	0.000
Var 4	0.661	0.000	0.000
Var 5	0.654	0.000	0.000
Var 6	0.652	0.000	0.000
Var 7	0.534	0.000	0.000
Var 8	0.500	0.000	0.000
Var 9	0.570	0.000	0.000
Var 10	0.522	0.000	0.000
Var 11	0.556	0.000	0.000
Var 12	0.543	0.000	0.000
Var 13	0.598	0.000	0.000
Var 14	0.633	0.000	0.000
Var 15	0.488	0.000	0.000
Var 16	0.625	0.000	0.000
Var 17	0.645	0.000	0.000
Var 18	0.631	0.000	0.000
Var 19	0.592	0.000	0.000
Var 20	0.814	0.000	0.000
Var 21	0.462	0.000	0.000
Var 22	0.506	0.000	0.000
Var 23	0.771	0.000	0.000
Var 24	0.519	0.000	0.000
Var 25	0.599	0.000	0.000
Var 26	0.599	0.000	0.000
Var 27	0.579	0.000	0.000
Var 28	0.665	0.000	0.000
Var 29	0.644	0.000	0.000
Var 30	0.000	0.399	0.000
Var 31	0.000	0.195	0.000
Var 32	0.000	0.480	0.000
Var 33	0.000	0.392	0.000
Var 34	0.000	0.443	0.000
Var 35	0.000	0.603	0.000
Var 36	0.000	0.425	0.000
Var 37	0.000	0.500	0.000
Var 38	0.000	0.452	0.000
Var 39	0.000	0.243	0.000
Var 40	0.000	0.433	0.000
Var 41	0.000	0.347	0.000
Var 42	0.000	0.378	0.000
Var 43	0.000	0.326	0.000
Var 44	0.000	0.411	0.000
Var 45	0.000	0.444	0.000
Var 46	0.000	0.516	0.000
Var 47	0.000	0.398	0.000
Var 48	0.000	0.448	0.000
Var 49	0.000	0.419	0.000
Var 50	0.000	0.538	0.000
Var 51	0.000	0.444	0.000
Var 52	0.000	0.462	0.000
Var 53	0.000	0.426	0.000
Var 54	0.000	0.463	0.000
Var 55	0.000	0.463	0.000
Var 56	0.000	0.595	0.000
Var 57	0.000	0.464	0.000
Var 58	0.000	0.489	0.000
Var 59	0.000	0.624	0.000
Var 60	0.000	0.510	0.000
Var 61	0.000	0.599	0.000
Var 62	0.000	0.000	0.266
Var 63	0.000	0.000	0.389
Var 64	0.000	0.000	0.440
Var 65	0.000	0.000	0.333
Var 66	0.000	0.000	0.535
Var 67	0.000	0.000	0.554
Var 68	0.000	0.000	0.393
Var 69	0.000	0.000	0.590
Var 70	0.000	0.000	0.330
Var 71	0.000	0.000	0.393
Var 72	0.000	0.000	0.518
Var 73	0.000	0.000	0.442
Var 74	0.000	0.000	0.667
Var 75	0.000	0.000	0.356
Var 76	0.000	0.000	0.367
Var 77	0.000	0.000	0.593
Var 78	0.000	0.000	0.475
Var 79	0.000	0.000	0.210
Var 80	0.000	0.000	0.479
Var 81	0.000	0.000	0.269
Var 82	0.000	0.000	0.446
Var 83	0.000	0.000	0.389
Var 84	0.000	0.000	0.508
Var 85	0.000	0.000	0.480
Var 86	0.000	0.000	0.447
Var 87	0.000	0.000	0.349
Var 88	0.000	0.000	0.435
Var 89	0.000	0.000	0.676
Var 90	0.000	0.000	0.417
Var 91	0.000	0.000	0.696
Var 92	0.000	0.000	0.573
Var 93	0.000	0.000	0.518
Var 94	0.000	0.000	0.619

Phi

	KSI 1	KSI 2	KSI 3
KSI 1	1.000		
KSI 2	0.676	1.000	
KSI 3	0.771	0.893	1.000

Total coefficient of determination for X-variables is 0.997.

Chi-square with 4274 degrees of freedom = 625285.97 (p = 0.000)
Goodness of Fit Index = 0.965
Adjusted Goodness of Fit Index = 0.963
Root Mean Square Residual = 0.040

Summary Statistics for Fitted Residuals
Smallest Fitted Residual = -0.197
Median Fitted Residual = -0.002
Largest Fitted Residual = 0.435

LISREL 7:
Estimation of Linear Structural Equation Systems
Program Version 7.16

Distributed by:
Scientific Software, Inc.
1369 Neitzel Road
Mooresville, Indiana 46158
(317) 831-6336

MVS - L I S R E L 7.16
By
Karl G. Joreskog and Dag Sorbom

The following LISREL control lines have been read:

```
DA NI=94 NO=43317 MA=PM
PM UNIT=18
DM UNIT=17
MO NX=94 NK=11 LX=FR TD=DI,FR PH=ST
PA LX

6*(1 0 0 0 0 0 0 0 0 0 0)
5*(0 1 0 0 0 0 0 0 0 0 0)
6*(0 0 1 0 0 0 0 0 0 0 0)
6*(0 0 0 1 0 0 0 0 0 0 0)
6*(0 0 0 0 1 0 0 0 0 0 0)
4*(0 0 0 0 0 1 0 0 0 0 0)
8*(0 0 0 0 0 0 1 0 0 0 0)
7*(0 0 0 0 0 0 0 1 0 0 0)
7*(0 0 0 0 0 0 0 0 1 0 0)
6*(0 0 0 0 0 0 0 0 0 1 0)
33*(0 0 0 0 0 0 0 0 0 0 1)
OU DWLS TM=1500
September 1989. Eleven factor solution.
```

Number of Input Variables	94
Number of Y-Variables	0
Number of X-Variables	94
Number of ETA-Variables	0
Number of KSI-Variables	11
Number of Observations	43317

Warning: Chi-square, standard errors, t-values and standardized residuals are calculated under the assumption of multi-variate normality.

BEST COPY AVAILABLE

LISREL Estimates (Diagonally Weighted Least Squares)
Lambda X

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6	KSI 7	KSI 8	KSI 9	KSI 10	KSI 11
Var 1	0.767	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 2	0.747	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 3	0.701	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 4	0.810	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 5	0.798	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 6	0.803	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 7	0.000	0.710	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 8	0.000	0.667	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 9	0.000	0.764	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 10	0.000	0.704	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 11	0.000	0.745	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 12	0.000	0.000	0.610	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 13	0.000	0.000	0.669	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 14	0.000	0.000	0.711	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 15	0.000	0.000	0.545	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 16	0.000	0.000	0.701	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 17	0.000	0.000	0.724	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 18	0.000	0.000	0.000	0.728	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 19	0.000	0.000	0.000	0.685	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 20	0.000	0.000	0.000	0.925	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 21	0.000	0.000	0.000	0.541	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 22	0.000	0.000	0.000	0.586	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 23	0.000	0.000	0.000	0.873	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Var 24	0.000	0.000	0.000	0.000	0.659	0.000	0.000	0.000	0.000	0.000	0.000
Var 25	0.000	0.000	0.000	0.000	0.752	0.000	0.000	0.000	0.000	0.000	0.000
Var 26	0.000	0.000	0.000	0.000	0.749	0.000	0.000	0.000	0.000	0.000	0.000
Var 27	0.000	0.000	0.000	0.000	0.723	0.000	0.000	0.000	0.000	0.000	0.000
Var 28	0.000	0.000	0.000	0.000	0.829	0.000	0.000	0.000	0.000	0.000	0.000
Var 29	0.000	0.000	0.000	0.000	0.802	0.000	0.000	0.000	0.000	0.000	0.000
Var 30	0.000	0.000	0.000	0.000	0.000	0.427	0.000	0.000	0.000	0.000	0.000
Var 31	0.000	0.000	0.000	0.000	0.000	0.210	0.000	0.000	0.000	0.000	0.000
Var 32	0.000	0.000	0.000	0.000	0.000	0.514	0.000	0.000	0.000	0.000	0.000
Var 33	0.000	0.000	0.000	0.000	0.000	0.419	0.000	0.000	0.000	0.000	0.000
Var 34	0.000	0.000	0.000	0.000	0.000	0.000	0.486	0.000	0.000	0.000	0.000
Var 35	0.000	0.000	0.000	0.000	0.000	0.000	0.663	0.000	0.000	0.000	0.000
Var 36	0.000	0.000	0.000	0.000	0.000	0.000	0.466	0.000	0.000	0.000	0.000
Var 37	0.000	0.000	0.000	0.000	0.000	0.000	0.549	0.000	0.000	0.000	0.000
Var 38	0.000	0.000	0.000	0.000	0.000	0.000	0.496	0.000	0.000	0.000	0.000
Var 39	0.000	0.000	0.000	0.000	0.000	0.000	0.267	0.000	0.000	0.000	0.000
Var 40	0.000	0.000	0.000	0.000	0.000	0.000	0.476	0.000	0.000	0.000	0.000
Var 41	0.000	0.000	0.000	0.000	0.000	0.000	0.381	0.000	0.000	0.000	0.000
Var 42	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.465	0.000	0.000	0.000
Var 43	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.399	0.000	0.000	0.000
Var 44	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.510	0.000	0.000	0.000
Var 45	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.551	0.000	0.000	0.000
Var 46	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.640	0.000	0.000	0.000
Var 47	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.495	0.000	0.000	0.000
Var 48	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.553	0.000	0.000	0.000
Var 49	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.448	0.000	0.000
Var 50	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.577	0.000	0.000
Var 51	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.476	0.000	0.000
Var 52	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.496	0.000	0.000
Var 53	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.456	0.000	0.000
Var 54	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.496	0.000	0.000
Var 55	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.497	0.000	0.000
Var 56	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.668	0.000
Var 57	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.521	0.000
Var 58	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.550	0.000
Var 59	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.703	0.000
Var 60	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.573	0.000
Var 61	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.664	0.000
Var 62	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.265
Var 63	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.389
Var 64	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.440
Var 65	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.333
Var 66	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.534
Var 67	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.554
Var 68	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.393

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6	KSI 7	KSI 8	KSI 9	KSI 10	KSI 11
Var 69	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.590
Var 70	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.330
Var 71	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.393
Var 72	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.518
Var 73	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.442
Var 74	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.667
Var 75	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.356
Var 76	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.367
Var 77	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.594
Var 78	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.475
Var 79	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.210
Var 80	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.480
Var 81	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.269
Var 82	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.446
Var 83	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.389
Var 84	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.508
Var 85	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.480
Var 86	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.447
Var 87	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.349
Var 88	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.435
Var 89	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.676
Var 90	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.417
Var 91	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.695
Var 92	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.573
Var 93	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.518
Var 94	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.619

Phi

	KSI 1	KSI 2	KSI 3	KSI 4	KSI 5	KSI 6	KSI 7	KSI 8	KSI 9	KSI 10	KSI 11
KSI 1	1.000										
KSI 2	0.562	1.000									
KSI 3	0.635	0.620	1.000								
KSI 4	0.539	0.570	0.696	1.000							
KSI 5	0.503	0.499	0.600	0.622	1.000						
KSI 6	0.492	0.435	0.581	0.531	0.447	1.000					
KSI 7	0.520	0.460	0.603	0.529	0.465	0.867	1.000				
KSI 8	0.428	0.374	0.492	0.438	0.351	0.806	0.754	1.000			
KSI 9	0.518	0.486	0.624	0.574	0.531	0.831	0.750	0.685	1.000		
KSI 10	0.451	0.429	0.562	0.560	0.507	0.825	0.706	0.627	0.791	1.000	
KSI 11	0.627	0.565	0.737	0.660	0.595	0.855	0.814	0.728	0.832	0.785	1.000

Total coefficient of determination for X-variables is 1.000.

Chi-square with 4222 degrees of freedom = 228892.07 (p = 0.000)
Goodness of Fit Index = 0.987
Adjusted Goodness of Fit Index = 0.986
Root Mean Square Residual = 0.027

Summary Statistics for Fitted Residuals
Smallest Fitted Residual = -0.255
Median Fitted Residual = 0.000
Largest Fitted Residual = 0.248