DOCUMENT RESUME

ED 469 172 TM 034 477

AUTHOR	Reese, Lynda M.
TITLE	A Classical Test Theory Perspective on LSAT Local Item Dependence. LSAC Research Report Series. Statistical Report.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-96-01
PUB DATE	1999-08-00
NOTE	16p.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; Item Response Theory; Law Schools; Test Construction; Test Items; *Test Theory
IDENTIFIERS	*Law School Admission Test; *Local Independence (Tests)

ABSTRACT

This study extended prior Law School Admission Council (LSAC) research related to the item response theory (IRT) local item independence assumption into the realm of classical test theory. Initially, results from the Law School Admission Test (LSAT) and two other tests were investigated to determine the approximate state of local item independence (LID) found in actual test data. Yen's Q3 statistic was used for this purpose. Based on these analyses, four levels of LID were defined, and associated data sets generated. The average case was defined to represent the LSAT. Values of the r-biserial statistic and the alpha reliability index were studied to determine the effect of LID on these measures. Percentile ranks were also studied in order to assess the impact of LID for individual test takers. The results indicated that for extreme cases of LID, the discrimination power of individual items and the reliability of the total test are overestimated. Percentile ranks were also clearly affected by the introduction of a high level of LID, indicating that the impact for individual test takers should be of concern. Because the LID became problematic only at the most extreme level stipulated, the less than extreme level of LID typically displayed by the LSAT is probably not a problem with respect to these particular outcomes. (Contains 9 tables and 10 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

A Classical Test Theory Perspective on LSAT Local Item Dependence

Lynda M. Reese
Law School Admission Council

Law School Admission Council
Statistical Report 96-01
August 1999

LAW Services

A Publication of the Law School Admission Council

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.

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

In the analysis of individual Law School Admission Test (LSAT) questions and the assembly and evaluation of LSAT forms, item response theory (IRT) is applied. IRT is a mathematical model that relates the probability that a test taker will answer a test question correctly to the ability level of the test taker. In applying IRT, a formal assumption of local item independence is made. This assumption states that, once the ability level of the test taker is accounted for, the responses of the test taker to individual items on the test should be statistically independent. In other words, test taker ability should be the only factor contributing to the test taker's performance on an item.

In a test taking situation, many circumstances arise that cause the local item independence assumption to be violated to some degree. For instance, in the Reading Comprehension section of the LSAT, several test questions are based on a common reading passage. Here, regardless of a test taker's reading comprehension ability, individual performance on all of the questions related to a particular passage may be enhanced or hindered by a prior level of knowledge of the subject matter of the passage. To the extent that a factor such as this causes a test taker to perform similarly on certain test questions, the test is said to exhibit some degree of local item dependence (LID).

Because local item independence is formally assumed in applications of IRT, LSAT research related to the local item independence assumption has centered primarily around IRT outcomes and measures. However, a category of statistics commonly known as classical statistics is also applied for certain purposes in the analysis of LSAT questions and test forms. Among the classical statistics applied in the analysis of LSAT data are an item/test correlation statistic (called the *r-biserial*), an index of reliability, and the percentile rank associated with each LSAT score. The *r-biserial* is an index of item discrimination used as an indicator of how well a test question distinguishes between more and less able test takers. This statistic is a common index of the quality of a test question. The reliability coefficient, calculated for each form of the LSAT, is an indicator of how consistent an individual's test score would be if they were to take the same form of the test many times. The percentile rank associated with a particular test score is the percent of test takers falling below that test score. Percentile ranks are routinely provided to test takers and law schools.

Although local item independence is not formally required for classical statistics, the impact of LID on these measures should be better understood. That is, to the extent that these statistics are influenced by the unknown effects of LID, the decisions made and procedures carried out using these statistics may be compromised. Therefore, this study extends past LSAT research related to the IRT local item independence assumption into the realm of classical test theory. Initially, results from the LSAT and two other tests were investigated to determine the approximate state of LID found in actual test data. Based on these analyses, four levels of LID (zero, low, medium, and high) were defined and associated data sets generated. Here, the medium LID level was defined to represent the LSAT. The classical statistics described above were studied in order to determine the effect of LID on these measures.

The results indicated that for extreme cases of LID, the discrimination power of individual items and the reliability of the total LSAT are overestimated. Percentile ranks were also clearly affected by the introduction of a high level of LID, indicating that the impact for individual test takers should be of concern. Because the LID became problematic only at the most extreme level simulated, the less than extreme level of LID typically displayed by the LSAT is probably not problematic with respect to these particular outcomes.

Abstract

This study extends past Law School Admission Council (LSAT) research related to the item response theory (IRT) local item independence assumption into the realm of classical test theory. Initially, results from the LSAT and two other tests were investigated to determine the approximate state of local item dependence (LID) found in actual test data. Yen's Q_3 statistic was employed for this purpose. Based on these analyses, four levels of LID were defined and associated data sets generated. Here, the average case was defined to represent the LSAT. Values of the *r-biserial* statistic and the α reliability index were studied to determine the effect of LID on these measures. Percentile ranks were also studied in order to assess the impact of LID for individual test takers. The results indicated that for extreme cases of LID, the discrimination power of individual items and the

reliability of the total test are overestimated. Percentile ranks were also clearly affected by the introduction of a high level of LID, indicating that the impact for individual test takers should be of concern. Because the LID became problematic only at the most extreme level simulated, the less than extreme level of LID typically displayed by the LSAT is probably not problematic with respect to these particular outcomes.

Introduction

Past research concerning violations of the local item independence assumption for the Law School Admission Council (LSAT) has centered primarily around item response theory (IRT) outcomes and measures. This is quite natural, given that the local item independence assumption is a formal requirement for IRT applications. However, classical statistics are still derived for the LSAT and utilized for some purposes. Statistics such as the item/test correlation (*r-biserial*) are useful tools for evaluating item quality, and many practitioners still use this statistic in the test assembly process. Although the *r-biserial* is not used in the assembly of LSAT forms, this statistic is studied in evaluating the quality of individual LSAT pretest items. Classical reliability indices are also useful for evaluating individual test forms, and a reliability index is routinely calculated for each form of the LSAT. In terms of the individual test taker, percentile ranks are often used to evaluate the standing of an individual with respect to other test takers, and such indices are reported to LSAT takers. Although local item independence is not formally assumed for these measures, the impact of local item dependence (LID) on these measures should be better understood. That is, to the extent that these statistics are influenced by the unknown effects of LID, the decisions made and procedures carried out using these statistics may be compromised. Therefore, the effect that LID has on these measures is still of interest to the practitioner. Given this concern, the goal of this study was to extend past LID research beyond the realm of IRT to determine the impact of this property for classical statistics. The results obtained should aid in the interpretation of these statistics for various purposes.

Local Item Dependence in Real Data

In order to realistically model LID, estimates of the true state of LID found within actual test data had to be obtained. Because it was desirable that these levels of LID be realistic, data from the LSAT were studied, along with data from the Pre-American College Test Plus (P-ACT+) and the Graduate Management Admission Test (GMAT). All three tests are similar in that they are large-scale high-stakes tests of acquired skills. The LSAT and GMAT are used as aids in graduate-level admissions decisions, while the P-ACT+ is administered to tenth-grade students. All of the tests have a large verbal component, and all contain a reading comprehension section. However, the GMAT and P-ACT+ add a quantitative dimension, and the P-ACT+ also includes a science measure.

The level of LID displayed by the real data was explored by employing Yen's (1984) Q_3 statistic. For two items i and j, the statistic is

$$Q_3 = r_{d_i d_j}, \tag{1}$$

a correlation among d_i and d_j values. For test taker k (adding an identifying subscript),

$$d_{ik} = u_{ik} - P_i(\theta_k), \tag{2}$$

where u_{ik} represents the score of the k^{th} test taker on item i (one if correct, zero otherwise) and $P_i(\theta_k)$ represents the probability of test taker k responding correctly to item i. The data for these three tests were calibrated using BILOG (386 BILOG 3, Mislevy & Bock, 1990) to obtain estimates of $P(\theta)$. The default scoring method, number of quadrature points, and priors for the item and ability parameters were used. All calibrations converged normally.

Q_3 values were calculated for each pair of items for each of the three tests studied. Summary statistics of the Q_3 statistics were evaluated within- and between-test sections and within- and between-item sets. The results of

these analyses (see Reese, 1995a) were then used to define the levels of LID to be simulated. It should be noted here that Yen (1993) has observed that Q_3 tends to have a slightly negative bias because IRT item probabilities that assume local item independence are used in its calculation. Therefore, she suggests that a Q_3 value of $-1/(n-1)$, where n represents the number of items in the test being analyzed, is expected when there is no LID. The deviation of the Q_3 statistic from this "criterion" value (Q_3 - criterion value) was used in the definition of the LID levels to be simulated in this study.

Tables 1 through 4 present the starting Q_3 values and the simulated Q_3 values for the zero, low, medium, and high LID levels, respectively. The cells of these tables represent the four sections of the test data to be simulated, with the diagonal elements representing the within-section LID and the off-diagonal elements representing the between-section LID. Since LSAT item parameters were used as the generating parameters, a four-section, 101-item test was simulated. The sections have 24, 24, 25, and 28 items; section sizes typically found in the LSAT. The starting values in these tables represent the Q_3 values specified in the data generation step, whereas the simulated values represent the mean Q_3 values actually recovered by the data generation process. Each of these cells will be discussed in greater detail below.

TABLE 1
Q_3 starting values and simulated recovered values for the zero LID level

	Section 1	Section 2	Section 3	Section 4
Section 1				
Starting	0.000			
Simulated	0.000			
Section 2				
Starting	0.000	0.000		
Simulated	0.001	0.001		
Section 3				
Starting	0.000	0.000	0.000	
Simulated	0.000	0.000	0.000	
Section 4				
Starting	0.000	0.000	0.000	0.000
Simulated	-.001	0.000	0.000	0.000

TABLE 2
Q_3 starting values and simulated recovered values for the low LID level

	Section 1	Section 2	Section 3	Section 4
Section 1				
Starting	0.010			
Simulated	0.012			
Section 2				
Starting	0.000	0.010		
Simulated	0.000	0.012		
Section 3				
Starting	0.000	0.010	0.010	
Simulated	0.000	0.012	0.010	
Section 4				
Starting	0.000	0.000	0.000	0.010
Simulated	0.000	0.001	0.000	0.011

TABLE 3
Q_3 starting values and simulated recovered values for the medium LID level

	Section 1	Section 2	Section 3	Section 4
Section 1				
Starting	0.050			
Simulated	0.046			
Section 2				
Starting	0.010	0.020		
Simulated	0.011	0.017		
Section 3				
Starting	0.010	0.020	0.020	
Simulated	0.012	0.017	0.016	
Section 4				
Starting	0.010	0.010	0.010	0.030
Simulated	0.014	0.009	0.010	0.027

TABLE 4
Q_3 starting values and simulated recovered values for the high LID level

	Section 1	Section 2	Section 3	Section 4
Section 1				
Starting	0.300			
Simulated	0.329			
Section 2				
Starting	0.050	0.300		
Simulated	0.053	0.308		
Section 3				
Starting	0.050	0.300	0.300	
Simulated	0.049	0.299	0.288	
Section 4				
Starting	0.050	0.050	0.050	0.300
Simulated	0.054	0.053	0.049	0.326

The starting values for simulating various levels of LID are presented as the top values in each cell of Tables 1 through 4. Data for the zero LID level, represented by Table 1, were simply generated by specifying zero LID between each pair of items. As displayed in Table 2, the low LID level was defined by assigning a low within-section starting value of .01 for all within-section cells. This value represents the lowest degree of within-section LID observed in the real data analyses rounded to the nearest significant digit. Note that for the low, medium, and high LID levels, the LID between sections 2 and 3 was assigned the same starting value as assigned within sections, because on the LSAT, these sections represent the same content, logical reasoning. For all other between-section cells, a value of 0.00, representing no LID, was assigned. Medium LID was defined by the values presented in Table 3. The within-section LID displayed by the LSAT was used here to define the within-section LID, because the LID displayed by this test was judged to be about average in comparison to the other two tests studied. The between-section value, for which a constant was chosen (with the exception of the between-section value for sections 2 and 3, as discussed above), was determined by studying the within-section LID displayed by each of the three tests analyzed. The value of .01 represented an approximately average level of between-section LID. Finally, Table 4 represents the starting values for the high LID level. The within-section LID was defined as the highest within-set LID observed, and the between-section LID was defined by the highest between-set LID observed.

Data Generation

To create simulated data, item responses (0 or 1) were generated to match the LID structures defined in Tables 1 through 4. Item parameter estimates obtained from a typical LSAT calibration were treated as true item parameters. Figure 1 overlays all of the item characteristic curves for this test and demonstrates the diversity of the item parameters being used as true parameters. For each level of LID defined, responses to a 101-item test consisting of four sections were simulated for 4,000 test takers with standard normally distributed ability values. The same sample of ability values was used in the data generation for each of the LID levels. The sample of generated ability values was rescaled by a linear transformation to ensure that the sample mean and standard deviation were exactly zero and one, respectively. Specifically, the steps followed in the data generation are as follows:

Step 1: A desired Q_3 correlation structure among the items was defined (see top row in each cell of Tables 1 through 4).

Step 2: The target correlations were adjusted by applying the equation

$$Q^*_{3ij} = 1.8 Q_{3ij}, \tag{3}$$

where Q_{3ij} may be found in the top row of each cell of Tables 1 through 4. The reason for this adjustment will be discussed further below.

Step 3: A vector x of multivariate random deviates was generated according to the adjusted Q_3 values derived in step 2.

Step 4: Using the inverse normal transformation, x was transformed to y, a vector of uniform (0,1) deviates. The adjustment applied in step 2 was necessary because the inverse normal transformation applied here is nonlinear. Results described below will demonstrate that the adjustment was quite effective.

Step 5: Each uniform deviate was compared to individual item correct probabilities in order to obtain 0 or 1 item responses.

For a more detailed description of the data generation method, see Pashley and Reese (1999).

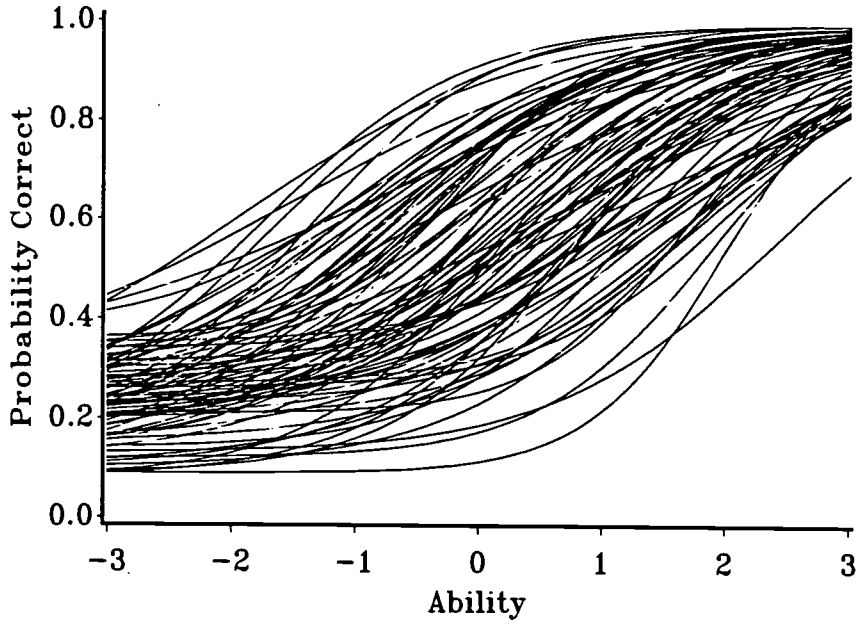

FIGURE 1. *Overlay plot of true item characteristic curves*

Accuracy of Data Simulation

At the outset, it was important to verify that the data generation had produced the desired LID structure. The lower values presented in each cell of Tables 1 through 4, labeled as "simulated," represent the levels of LID achieved in the simulated data. These values were calculated by utilizing the true item and ability parameters and the generated item responses. Comparing these values to those directly above them in each cell of these tables reveals the high degree of accuracy achieved by the data generation method. The zero LID level was achieved almost exactly. The low LID level was also achieved very well, with discrepancies found only in the third decimal place. Differences of this degree are not considered problematic. For the medium LID level, all local dependencies achieved round to the desired values. Finally, for the high LID level, the within-section LID of .3 and the between-section LID of .05 were both achieved quite well.

The Effects of Local Item Dependence on Classical Statistics

The Effect of LID on the R-biserial Statistic

In the analyses carried out to determine the effect of LID on the *r-biserial* statistic, the zero LID level, which represents the case of perfect local item independence, was used as the reference for comparison. That is, the *r-biserial* statistic calculated for the low, medium, and high LID levels was compared to those calculated for the zero LID level.

Tables 5 and 6 present summary statistics and correlation coefficients for the *r-biserial* statistic for each LID level. These were inspected visually in order to gain a general sense of where differences arise. The summary statistics for the *r-biserial* presented in Table 5 indicate that the mean value of this statistic increases as the LID increases. The change in the mean for this statistic is very slight as we go from the zero LID level to the low and medium LID levels. The high LID level shows the greatest change. The standard deviation remains fairly constant, however, across all four LID levels. The correlation coefficients for this statistic, presented in Table 6, are consistent with these results, as the *r-biserials* for the low and medium LID levels are very strongly correlated with the *r-biserials* for the zero LID level (.97 and .95, respectively), but the correlation between the zero and high LID levels drops to .72.

TABLE 5
Summary statistics for the r-biserial by LID level

	Zero	Low	Medium	High
Mean	0.42	0.43	0.45	0.63
S.D.	0.09	0.09	0.08	0.09

TABLE 6
Pearson correlation coefficients among r-biserial values by LID level

	Low	Medium	High
Zero	0.97	0.95	0.72
Low		0.96	0.72
Medium			0.65

The root mean squared error (RMSE) statistic was calculated to compare the values of the *r-biserial* statistics for each of the dependent LID levels (low, medium, and high) to the zero LID level. This was accomplished by calculating

$$RMSE = [\frac{1}{n}\sum_{i=1}^{n}(I - D)^2]^{1/2}, \quad (4)$$

where n represents the number of items, I represents the independent *r-biserial* statistic (zero LID level), and D represents the *r-biserial* for the low, medium, or high LID level. The *bias* statistic comparing the *r-biserial* values was calculated as follows

$$Bias = \frac{1}{n}\sum_{i=1}^{n}(I - D). \quad (5)$$

Table 7 presents the *bias* and RMSE statistics for the *r-biserial*. The results presented here are consistent with results discussed earlier. The overestimation of the *r-biserial* is clearly borne out in this analysis. A slight negative *bias* at the low and medium LID levels becomes much stronger for the high LID level. Similarly, the error in this statistic increases as the LID increases.

TABLE 7
RMSE and bias statistic between the independent and dependent r-biserial *values*

	LID Level		
	Low	Medium	High
Bias	-.01	-.03	-.21
RMSE	0.02	0.04	0.22

The results presented here for the classical index of discrimination are consistent with what has been observed for IRT statistics. Here, analyses of the zero through high LID data revealed that the *a*-parameter is overestimated for the high LID level (Reese, 1995b). Masters (1988) and Yen (1993) have also noted that an increase in LID causes the *a*-parameter to be overestimated. They attribute this overestimation to the fact that a

strengthening of the relationship between two items causes the relationship between an item and the total test to be strengthened, thereby causing the item to appear to be more discriminating than it really is. It seems logical that this same effect would be observed for both classical and IRT indices of item discrimination. For the LSAT, these results are very positive in that the *r-biserial* statistic was not seriously affected by the degree of LID defined for the medium LID level. This result indicates that this statistic is not adversely affected by the degree of LID generally present for LSAT data.

The Effect of LID on Reliability

A coefficient α reliability index (Nunnally, 1978) was calculated for each of the simulated datasets; and these values are presented in the first column of Table 8. This statistic is fairly high for the zero LID level (0.91), as would be expected for a 101-item test. Reliability values for the low and medium LID levels represent only a slight increase over the zero LID value, but the reliability index for the high LID level is more extreme. To clarify the magnitude of these differences, a derivation of the Spearman-Brown prophesy formula (Nunnally, 1978, p. 211) was applied to determine the extent to which a test of zero LID would have to be lengthened in order to achieve each of the three dependent reliability indexes. The equation applied in this analysis is given by

$$n_{new} = \frac{r_D - (r_D r_I)}{r_I - (r_D r_I)} n_{old}, \tag{6}$$

where n_{new} is the new test length, n_{old} is the length of the zero LID test (i.e., 101 items), r_D is the reliability coefficient calculated for the dependent data (low, medium, or high LID), and r_I is the reliability coefficient calculated for the zero LID data. This analysis revealed that the zero LID test would have to be lengthened by 14 items to achieve the reliability index that was observed for the low LID level. Thirty-two additional items would be needed to achieve the reliability index that was observed for the medium LID level. In sharp contrast, the length of the zero LID test would have to be more than tripled in order to achieve the reliability index of .97 estimated for the high LID data. On the surface, these results might suggest that LID is actually good for a test as it seems to increase reliability. However, these results are deceiving.

TABLE 8
Analysis of the coefficient α and $\hat{\rho}$ by LID level

	α	Number of items needed for zero LID test of this α	$\hat{\sigma}_T^2$	$\hat{\sigma}_x^2$	$\hat{\sigma}_e^2$	$\hat{\rho}$	Number of items needed for zero LID test of this ρ
Zero	0.91	101	211.6	230.3	18.7	0.92	101
Low	0.92	115	211.6	239.2	27.6	0.88	64
Medium	0.93	133	211.6	264.9	53.3	0.80	35
High	0.97	323	211.6	518.3	306.7	0.41	6

The remainder of Table 8 presents a true-score analysis of these effects. Using the generating item and ability parameters, an estimate of the true-score variance ($\hat{\sigma}_T^2$) was derived. The observed-score variance ($\hat{\sigma}_x^2$) was simply calculated from the observed-score data for each LID level. Using these values, an estimate of the true reliability ($\hat{\rho}$) was derived by applying the equation

$$\hat{\rho} = \frac{\hat{\sigma}_T^2}{\hat{\sigma}_X^2}. \tag{7}$$

In studying these values, found in column 6 of Table 8, it is clear that the LID modeled here actually reduces the true reliability of the test, as would be expected.

The final column of Table 8 confirms the true influence of LID on reliability. Here, the Spearman-Brown prophesy formula was again applied to determine the effective test length for each of the levels of LID. We see that a 101-item test of low LID is actually comparable to a 64-item test of zero LID and a 101-item test of medium LID is comparable to a zero LID test of 35 items. Again, the results for the high LID level are the most alarming. A 101-item test of high LID is actually comparable to a zero LID test of only 6 items. Clearly, LID does not add to the precision of the test, but rather distracts from it.

Returning to the results obtained for the coefficient α, it is clear that the degree of LID displayed by the low and medium LID data causes some overestimation of this reliability index, whereas the LID simulated for the high LID level results in a strong overestimation of the reliability index. This overestimation is a direct result of the systematic error introduced into the data in the form of LID. In calculating coefficient α, split-halves of the test were analyzed, and the usual assumption was made that the errors on these split-halves were uncorrelated; a clearly incorrect assumption. Stronger LID results in more systematic error, and thus a stronger overestimation of the reliability coefficient. The true-score analyses presented in Table 8 demonstrate that not only can coefficient α overestimate the reliability of a test, as was reported by Sireci, Thissen, and Wainer (1991) and Wainer and Thissen (1996), but LID actually reduces the reliability of a test.

The IRT analog to the reliability coefficient is the test information function. Research has shown that LID causes this function to be overestimated (Reese, 1995b; Yen, 1993). So again, the results observed here for the classical statistics are consistent with what has been observed for IRT.

The Effect of LID on Percentile Ranks

Because it was of interest to determine the impact of LID for individual test takers, percentile ranks were studied. Percentile ranks for the LSAT are reported to test takers along with their individual LSAT scores, so the accuracy of this measure is certainly of interest. Initially, the percentile rank of each individual test taker was determined for each level of LID. Analyses were then carried out in order to compare the percentile ranks observed for the low, medium, and high LID levels to those observed for the zero LID level.

The RMSE and *bias* statistics were calculated to compare the values of the percentile ranks for each of the dependent LID levels (low, medium, and high) to the zero LID level. This was accomplished by applying Equations 4 and 5, where n represents the number of test takers, I represents the independent percentile ranks (zero LID level), and D represents the percentile ranks for the low, medium, or high LID levels.

The top row of Table 9 presents the results of this analysis. For the *bias* statistic, a positive value is observed for the low LID level, while a negative value is observed for the medium and high LID levels. These values are negligible in all instances. The RMSE statistic, however, more than doubles as the LID level increases from low to high. The reason for the disparity between these two statistics is borne out in the next set of analyses.

TABLE 9
RMSE and Bias statistic between the independent and dependent percentile ranks calculated for the total sample by quartile

	LID Level		
	Low	Medium	High
Total sample percentile rank			
Bias	0.04	-.06	-.01
RMSE	12.14	14.84	24.92
Independent percentile rank from 0 to 25			
Bias	-3.63	-5.32	-14.56
RMSE	11.49	14.18	26.50
Independent percentile rank from 26 to 50			
Bias	-1.18	-1.78	-4.54
RMSE	13.99	16.71	24.73
Independent percentile rank from 51 to 75			
Bias	2.43	2.60	6.05
RMSE	13.90	16.52	24.45
Independent percentile rank from 76 to 100			
Bias	2.60	4.41	13.48
RMSE	8.36	11.36	23.81

In order to determine if the impact of the LID varied according to the ability level of the test taker, test takers were grouped into four quartiles based on their percentiles for the zero LID level. The RMSE and *bias* statistics were then calculated again based on these groupings in order to determined if test takers at a particular level of ability are disadvantaged to a greater or lesser extent as a result of the LID.

The remaining portion of Table 9 presents the results of the analyses for the grouped percentile ranks. Here, the *bias* statistic indicates a tendency for the percentile ranks to be overestimated for the two lower quartiles and underestimated for the two higher quartiles. The bias for all quartiles starts off fairly low for the low LID level, increases for the medium LID level, and more than doubles for the high LID level. The extreme quartiles show the greatest effect due to the increasing LID. The degree of bias is similar for these two extreme groups, though opposite in direction. It seems clear that these values canceled each other out in the analyses run for the total group.

In terms of the RMSE statistic, the results indicate that the error increases as the LID increases for all quartiles. As has been the case for all other outcome measures evaluated in this research, a slight increase in this statistic occurs between the low and medium LID conditions, and a much greater increase occurs between the medium and high LID conditions. The error essentially doubles from the low to high LID level for the first three quartiles, and it nearly triples for the highest quartile as the LID is increased from low to high. The greatest error is incurred for the lowest quartile at the high LID level, but the RMSE values at this level are quite high and fairly similar for all four groups. For all three LID levels, RMSE values are a bit lower for the extreme percentile groups as compared to the two middle percentile groups. In fact, for all three levels of LID, the lowest RMSE values are observed for the highest quartile.

Previous research has reported that a high degree of LID causes low scores to be underestimated and high scores to be overestimated (Reese, 1995b). The results presented here indicate that at the high LID level, the underestimation of low scores results in higher percentile ranks for lower scoring test takers and the overestimation of high scores results in lower percentile ranks for higher scoring test takers. Therefore, when actual test scores are of interest, high LID results in an unfair advantage for high scoring test takers and an

unfair disadvantage for low scoring test takers. However, the opposite is true when percentile ranks are considered.

With respect to the LSAT, again, there is not much cause for concern. The introduction of medium LID did not have an extreme effect on the percentile ranks of individual test takers. To the extent that the LID observed for the LSAT remains constant across test forms, the percentile ranks of individual test takers are not adversely affected by this characteristic of the test data.

Discussion

The results obtained in this research should be of interest to those who rely on classical item and test statistics for the LSAT and for other tests. Fortunately, at the level of LID displayed by the LSAT, the effects are minimal and not problematic. However, an item presented in a high LID test or test section will appear to be more discriminating than is actually the case. Therefore, any decisions made concerning the quality of that item or its contribution to a test may be invalid. In terms of test assembly, a test assembled based on statistics derived within a test form or section displaying a high level of LID will not possess the level of discrimination or reliability predicted by the classical statistics. A better understanding of these issues should prove to be very valuable to those carrying out such procedures.

In terms of the results presented for the percentile ranks, the implications for individual test takers are clear. A high level of LID results in an unfair bias for this index of test taker performance. Here, the high scoring test taker is at a disadvantage while the performance of the low scoring test taker is overestimated. Because most admissions decisions are likely made at the middle and high scoring levels, the implications for the low scoring test takers are probably not as problematic. For LSAT takers, again, because no significant effects on this index were observed at the medium LID level, there is no cause for concern. The results presented here do, however, indicate that the effects of LID should be monitored for classical test theory outcomes as well as IRT outcomes for any testing program utilizing these measures.

With respect to item-level statistics, this study did not address the effect of LID on the estimation of item difficulty. In the data simulation routine applied here, the value of the $p+$ statistic, or the percent of test takers answering an item correctly, was held constant. Therefore, it was not possible to assess the effect of LID on this statistic with the data sets analyzed here. Previous research has indicated that high LID causes low scores to be underestimated and high scores to be overestimated (Reese, 1995b). It seems to logically follow that the effect of high LID on the $p+$ statistic would differ based on the ability level of the test takers. The $p+$ would be underestimated for low scoring test takers and overestimated for high scoring test takers. Such effects may wash out when an entire group of test takers is considered, resulting in minimal overall impact for this statistic. Further research should be carried out to verify this hypothesis.

References

Masters, G. N. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement, 25,* 15-29.

Mislevy, R. J., & Bock, R. D. (1990). *BILOG.* Mooresville, IN: Scientific Software, Inc.

Nunnally, J. C. (1978). *Psychometric Theory.* New York: McGraw-Hill.

Pashley, P. J., & Reese, L. M. (1999). *On generating locally dependent item responses* (Statistical Report 95-04). Newtown, PA: Law School Admission Council.

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.

Sireci, S. G., Thissen, D., & Wainer, H. (1991). On the reliability of testlet-based tests. *Journal of Educational Measurement*, *28*, 237-247.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*, 22-29.

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.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-214.

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").