ED 469 244 TM 034 466

AUTHOR	Reese, Lynda M.
TITLE	The Impact of Local Dependencies on Some LSAT Outcomes. Statistical Report. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-R-95-02
PUB DATE	1995-04-00
NOTE	29p.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC02 Plus Postage.
DESCRIPTORS	College Applicants; *College Entrance Examinations; Higher Education; Item Response Theory; *Law Schools; *Test Items
IDENTIFIERS	Calibration; Item Dependence; *Local Independence (Tests)

ABSTRACT

This study explored the impact of various degrees of violations of the item response theory (IRT) local independence assumption on the Law School Admission Test (LSAT) calibration and score distribution estimates. 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 Q3 statistic (W. Yen, 1984) was used for this purpose. Based on these analyses, four levels of LID were defined and associated data sets generated. Estimates from the simulated data were compared to their corresponding LSAT generating values in order to analyze the effects of the LID on IRT calibration and score distribution estimates. Results indicate that LID causes low scores to be underestimated and high scores to be overestimated. Effects on item and ability parameter and test characteristic curve estimation were clearly demonstrated. Score distributions were also markedly changed by the introduction of LID. The effects observed were mainly problematic for high LID levels. Deficiencies in Yen's Q3 statistic were also observed. (Contains 10 figures, 16 tables, and 32 references.) (Author/SLD)

LSAC RESEARCH REPORT SERIES

TM

The Impact of Local Dependencies on Some LSAT Outcomes

Lynda M. Reese

Law School Admission Council
Statistical Report 95-02
April 1995

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 193 law schools in the United States and Canada.

For information, write: Communications, Law School Admission Council, Box 40, 661 Penn Street, Newtown, PA 18940.

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

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

Table of Contents

Abstract iv

Introduction 1

Local Item Dependence in Real Data 2

Data Generation 4

Evaluation of Simulated Data 5

Accuracy of Data Simulation 5
Recovery of Local Item Dependence in Calibrated Data 6

The Effects of Local Item Dependence on Calibration Results 9

Analysis of Item and Ability Parameters 9
Analysis of the Item and Test Characteristic Curves 11

The Effect of Local Item Dependence on Scores 16

Score Distributions 16
Rank Order 17

Conclusions and Future Directions 18

Impact of Local Item Dependence on Calibration Results 18
Impact of Local Item Dependence on the Score Distribution 18
The Q3 Statistic 18
Implications for Measurement 19
Future Directions 20

References 21

Tables

Table 1 Q_3 Values to be Simulated for the Zero LID Level 3

Table 2 Q_3 Values to be Simulated for the Low LID Level 3

Table 3 Q_3 Values to be Simulated for the Medium LID Level 3

Table 4 Q_3 Values to be Simulated for the High LID Level 3

Table 5 Dependence Levels Recovered, Using True Item and Ability Parameters, from Zero LID Simulated Data 5

Table 6 Dependence Levels Recovered, Using True Item and Ability Parameters, from Low LID Simulated Data 5

Table 7 Dependence Levels Recovered, Using True Item and Ability Parameters, from Medium LID Simulated Data 6

Table 8 Dependence Levels Recovered, Using True Item and Ability Parameters, from High LID Simulated Data 6

Table 9 Within- and Between-Section Summary of Q_3 Statistics Using Estimated Item and Ability Parameters for the Zero LID Simulated Data 7

Table 10 Within- and Between-Section Summary of Q_3 Statistics Using Estimated Item and Ability Parameters for the Low LID Simulated Data 7

Table 11 Within- and Between-Section Summary of Q_3 Statistics Using Estimated Item and Ability Parameters for the Medium LID Simulated Data 8

Table 12 Within- and Between-Section Summary of Q_3 Statistics Using Estimated Item and Ability Parameters for the High LID Simulated Data 8

Table 13 Summary Statistics for Item Parameter Estimates 9

Table 14 Pearson Correlation Coefficients among Item Parameters 10

Table 15 Pearson Correlation Coefficients among Ability Values 11

Table 16 Spearman Rank Order Correlations 17

Figures

Figure 1 Overlay plot of true item characteristic curves 4

Figure 2 Item characteristic curve overlay plots for item 5 that exhibited a mild effect 12

Figure 3 Item characteristic curve overlay plots for item 24 that exhibited an average effect 13

Figure 4 Item characteristic curve overlay plots for item 33 that exhibited a strong effect 14

Figure 5 Test characteristic curve overlay plot 15

Figure 6 Frequency distribution overlay plot for the zero LID level 16

Figure 7 Frequency distribution overlay plot for the low LID level 16

Figure 8 Frequency distribution overlay plot for the medium LID level 16

Figure 9 Frequency distribution overlay plot for the high LID level 16

Figure 10 Test information function overlay plot 19

Abstract

This study explored the impact of various degrees of violations of the item response theory (IRT)-local independence assumption on the Law School Admission Test (LSAT) calibration and score distribution estimates. 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. Estimates from the simulated data were compared to their corresponding LSAT generating values in order to analyze the effects of the LID on IRT calibration and score distribution estimates.

The results indicated that LID causes low scores to be underestimated and high scores to be overestimated. Effects on item and ability parameter and test characteristic curve estimation were clearly demonstrated. Score distributions were also markedly changed by the introduction of LID. The effects observed were mainly problematic for high LID levels. Deficiencies in Yen's Q_3 statistic were also observed.

Introduction

For the purpose of equating new forms of the Law School Admission Test (LSAT) to previous forms, performing item analyses, and assembling new test forms, Law School Admission Council (LSAC) employs the three-parameter logistic (3PL)-item response theory (IRT) model. Here, the probability that a test taker will correctly answer a particular item is defined by

$$P_i(\theta) = c_i + (1-c_i)\{1 + \exp[-Da_i(\theta-b_i]\}^{-1}, \quad (1)$$

where, a_i, b_i, and c_i represent item i's discrimination, difficulty and guessing parameters, respectively; θ represents the ability level of the test taker; and D is a scaling factor usually set to 1.7. To facilitate the estimation of IRT parameters, an assumption of local item independence is usually made. This assumption states that the responses of test takers to individual items on a test must be statistically independent after conditioning on their ability levels. The local item independence assumption may be defined by the equation

$$P_{ij}(\theta) = \mathrm{P_i}(\theta)P_j(\theta), \quad (2)$$

which states that the probability of observing a pair of correct responses to two items, i and j, is the product of the individual correct item response probabilities. This equation holds only if the individual item responses are statistically independent, given test takers' ability levels.

More than a decade ago, Goldstein (1980, p. 239) stated that "there seems to have been little systematic attempt to carry out suitable experiments or to study the consequences for estimation and inference procedures when [the local item independence assumption] is violated. Without the results of such studies it is difficult to be sure how serious might be any failure of [the local item independence assumption]." A review of the current research on the local item independence assumption would indicate that this statement is no longer true. Many approaches have been taken in studying the issue of local item dependence (LID). Some researchers have identified the situations in which LID is likely to occur (Yen, 1993), while others have identified some of the consequences of LID for IRT (Masters, 1988; Yen, 1993). Some researchers have attempted to build models to account for the LID so that it might be allowed to occur (Ackerman, 1987; Ackerman & Spray, 1987; Andrich, 1978, 1985; Bell, Pattison, & Withers, 1988; Embretson, 1984; Jannarone, 1986, 1987, 1991a, 1991b, 1991c, in press; Kempf, 1977; Rosenbaum, 1988; Spray & Ackerman, 1987; Wainer & Kiely, 1987), while others have developed statistics for detecting LID in order that it might be avoided (Kelderman, 1984; Lord, 1953; Van den Wollenberg, 1982; Yen, 1984). Some researchers have approached the problem by applying their analyses to real data, while others have used a data-generation framework in studying this problem.

While much has been done toward addressing LID, there are still issues that remain to be investigated. Many researchers have attempted to develop measurement models which account for the LID found within test data; however, most of these methods are not suitable for practical use. The application of item bundles and testlets within an operational testing program seems more practical (Rosenbaum, 1988; Wainer & Kiely, 1987), but these solutions are specific to only one cause of LID. Also, the application of testlets could lead to failure of the parameter invariance assumption in that the parameters may not be invariant with respect to the specific items included in the testlet.

Stout (1987, 1990) suggested that in practice, the local item independence assumption could be weakened. Stout proposed a monotone nonparametric IRT framework in which the assumption of local item independence would be replaced with an assumption of "essential independence." In order to meet the essential independence requirement, the covariances between items (conditional on ability) must be small on average. Thus, rather than attempting to meet the strong requirement that the responses to test items be conditionally unrelated, which is probably impossible to meet in most practical situations, it would be sufficient to assure that the association between items is sufficiently weak. Other researchers (Holland, 1981; Rosenbaum, 1984) have also proposed a weakening of the local item independence assumption.

Given that most researchers would probably agree with the assertions of Stout, Holland, and Rosenbaum, the task of determining the effects of various levels of LID remains to be accomplished. The current study conducts analyses to determine the effect of various levels of LID on IRT calibration and the score distribution estimates for the LSAT.

The study was approached from both a real-data analysis and a data-generation framework, in that real data were first analyzed to determine the levels of LID that were likely to arise in actual testing situations. These levels of LID were used as a baseline in defining the levels of LID to be simulated in the data-generation phase of the study. The measurement statistics identified as susceptible to the effects of LID were then studied to determine how these outcomes were affected by the levels of LID simulated.

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. Since 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 10th-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 data for these three tests were calibrated using BILOG (Mislevy & Bock, 1990), with default priors for the item and ability parameters.

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

$$Q_3 = r_{d_i d_j} \tag{3}$$

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

$$d_{ik}=u_{ik}-P_i(\theta_k), \tag{4}$$

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.

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, 1995) 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 due to the fact that 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 value, referred to here as the "criterion" value, was used in the definition of the LID levels to be simulated in this study.

Tables 1 through 4 present the starting 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 average within-section LID and the off-diagonal elements representing the average 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 sections.

TABLE 1
Q3 Values to be Simulated for the Zero LID Level

Section	1	2	3	4
1	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00
3	0.00	0.00	0.00	0.00
4	0.00	0.00	0.00	0.00

TABLE 2
Q3 Values to be Simulated for the Low LID Level

Section	1	2	3	4
1	0.01	0.00	0.00	0.00
2	0.00	0.01	0.01	0.00
3	0.00	0.01	0.01	0.00
4	0.00	0.00	0.00	0.01

TABLE 3
Q3 Values to be Simulated for the Medium LID Level

Section	1	2	3	4
1	0.05	0.01	0.01	0.01
2	0.01	0.02	0.02	0.01
3	0.01	0.02	0.02	0.01
4	0.01	0.01	0.01	0.03

TABLE 4
Q3 Values to be Simulated for the High LID Level

Section	1	2	3	4
1	0.3	0.05	0.05	0.05
2	0.05	0.3	0.3	0.05
3	0.05	0.3	0.3	0.05
4	0.05	0.05	0.05	0.3

The zero LID level, represented by Table 1, was generated by simulating 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 level as assigned within sections, since 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. The between-section value, for which a constant was chosen, 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 (for further details, see Pashley & Reese, 1995). 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.

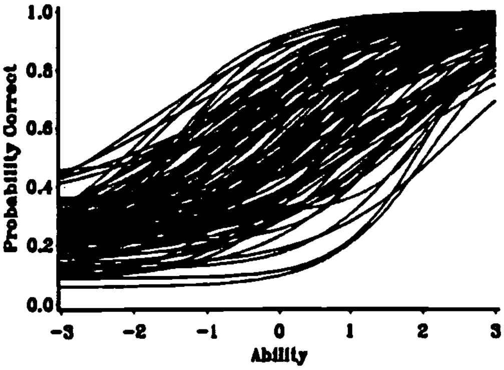


FIGURE 1. *Overlay plot of true item characteristic curves*

The simulated data were calibrated using BILOG (Mislevy & Bock, 1990). To assure that the item and ability parameter estimates for the true and dependent data were on a common scale, the ability parameters from each calibration were transformed to have a mean of zero and a standard deviation of one. (**Note:** The results for BILOG were again standardized in this way to ensure that all samples of ability estimates had means and standard deviations exactly equal to zero and one, respectively.) An associated transformation was applied to the *a*- and *b*-parameters as described by Hambleton and Swaminathan (1985, p. 126).

Evaluation of Simulated Data

Before any observations could be made based on analyses of the simulated data, it was necessary to verify that the data simulation method had produced data sets with the intended properties. This evaluation was made on two levels. First, the accuracy of the data simulation was evaluated. Next, analyses were carried out to determine the extent to which the true LID defined for the simulated data was recovered by the Q_3 statistic.

Accuracy of Data Simulation

At the outset, it was important to verify that the data generation had produced the desired LID structure. Tables 5 through 8 represent the true levels of LID achieved in the simulated data. These values were calculated by utilizing the true item and ability parameters and the simulated item responses. Comparing these values to those defined in Tables 1 through 4 reveals the high degree to which the LID was recovered in the simulated data. The zero LID level was recovered almost exactly. The low LID level was also recovered 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 dependencies recovered round to the desired values. Finally, for the high LID level, the within-section dependence of .3 and the between-section dependence of .05 were both recovered quite well.

TABLE 5
Dependence Levels Recovered, Using True Item and Ability Parameters, from Zero LID Simulated Data

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

TABLE 6
Dependence Levels Recovered, Using True Item and Ability Parameters, from Low LID Simulated Data

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

TABLE 7
Dependence Levels Recovered, Using True Item and Ability Parameters, from Medium LID Simulated Data

Section	1	2	3	4
1	0.046	0.011	0.012	0.014
2	0.011	0.017	0.017	0.009
3	0.012	0.017	0.016	0.010
4	0.014	0.009	0.010	0.027

TABLE 8
Dependence Levels Recovered, Using True Item and Ability Parameters, from High LID Simulated Data

Section	1	2	3	4
1	0.329	0.053	0.049	0.054
2	0.053	0.308	0.299	0.053
3	0.049	0.299	0.288	0.049
4	0.054	0.053	0.049	0.326

Recovery of Local Item Dependence in Calibrated Data

After the data were calibrated and scaled, Q_3 values were studied again to determine the extent to which this statistic could recapture the various LID levels. Tables 9 through 12 present the results of this analysis. Recall that in calculating the Q3 statistic from the estimated parameters, there is a tendency for the statistic to have a slightly negative bias due to the fact that IRT item probabilities that assume local item independence are used in its calculation. Therefore, a Q_3 value of $-1/(n-1)$, or -.01, represents zero LID for a 101-item test (Yen, 1993). This criterion value is presented in parentheses in Tables 9 through 12. Also in these tables, the Q_3 values observed within sections are presented on the diagonal, with the deviation from the criterion value presented in parentheses. The off-diagonal elements represent the between-section Q_3 values, with the deviation of these values from the criterion value presented in parentheses.

TABLE 9
Within- and Between-Section Summary of Q_3 Statistics Using Estimated Item and Ability Parameters for the Zero LID Simulated Data

Section	1 (-.01)	2 (-.01)	3 (-.01)	4 (-.01)
1	-.005 (0.005)			
2	-.006 (0.004)	-0.009 (0.001)		
3	-.006 (0.004)	.009 (0.001)	.009 (0.001)	
4	-.008 (0.002)	-.010 (0.000)	-.010 (0.000)	-.010 (0.000)

TABLE 10
Within- and Between-Section Summary of Q_3 Statistics
Using Estimated Item and Ability Parameters for the Low LID Simulated Data

Section	1 (-.01)	2 (-.01)	3 (-.01)	4 (-.01)
1	0.006 (0.016)			
2	-.010 (0.000)	-.005 (0.005)		
3	-0.10 (0.000)	-0.004 (0.006)	-0.004 (0.006)	
4	-.009 (0.001)	-.014 (-.004)	-.014 (-.004)	-.002 (0.008)

For the zero LID level presented in Table 9, the Q_3 statistic recaptured the LID quite well. A deviation from the criterion value of .005 is the highest observed within sections, while a value of .004 is the highest observed between sections. For the low LID level, Table 10 shows that all within-section deviations round to .01, with the exception of section 1 which rounds to .02. This is very encouraging. Between-section LIDs are all very close to zero, with some slightly negative values.

TABLE 11
Within- and Between-Section Summary of Q3 Statistics Using Estimated Item and Ability Parameters for the Medium LID Simulated Data

Section	1 (-.01)	2 (-.01)	3 (-.01)	4 (-.01)
1	0.021 (0.031)			
2	-.013 (-.003)	-.004 (0.006)		
3	-.012 (-.002)	-.004 (0.006)	-.004 (0.006)	
4	-.014 (-.004)	-.014 (-.004)	-.014 (-.004)	0.001 (0.011)

TABLE 12
Within- and Between-Section Summary of Q3 Statistics Using Estimated Item and Ability Parameters for the High LID Simulated Data

Section	1 (-.01)	2 (-.01)	3 (-.01)	4 (-.01)
1	0.294 (0.304)			
2	-.073 (-.063)	0.038 (0.048)		
3	-.074 (-.064)	0.039 (-.049)	0.039 (0.049)	
4	0.002 (0.012)	-.109 (-.099)	-.107 (-.097)	0.286 (0.296)

The medium LID results that are presented in Table 11 are not quite as encouraging as those observed for the zero and low LID. Recall that the LID levels defined here were intended to emulate the LSAT. The within-section values are all a bit lower than what was built into the simulated data. For the between-section values, the Q3 statistic indicates negative LID. The results for the high LID level, found in Table 12, are similar to the medium LID level. For sections 1 and 4, the within-section LID was recaptured by the Q3 statistic, but the LID within sections 2 and 3, while higher than that observed for any other LID levels, falls short of what was simulated. The between-section LID, while fairly strong, is displayed as negative LID in all but one case.

The results presented here for the simulated medium and high LID data were somewhat troublesome and may call into question the usefulness of the Q3 statistic for monitoring LID. The LID simulated for the medium LID level was that observed for an average test. These results indicate that the Q3 statistic underestimates LID at this level and at higher levels. For the high LID level, it is clear that the true LID is underestimated when a large block of items is being considered. Here, for section 1 and section 4, consisting of 24 and 28 items, respectively, the high LID level is recovered. It is for the between-section LID and the LID within and between sections 2 and 3 that the LID is underestimated. This result indicates that for a long string of dependent items, the Q_3 statistic underestimates the actual LID in the data. Perhaps Yen's correction factor of -1/(n-1), while appropriate for zero and low LID, is too conservative for average and high LID levels, especially for large blocks of dependent items.

The Effects of Local Item Dependence on Calibration Results

One goal of this research was to determine the effect of the various levels of LID on the calibration results. This evaluation was made by first studying the effect of LID on the item and ability parameters. As the item parameters combine to define the item characteristic curve, the effect on this variable was studied next, along with the effect on the test characteristic curve.

Analysis of Item and Ability Parameters

Table 13 presents the summary statistics of the true and estimated IRT item parameters for each level of LID. The ability parameter θ, is not included in this table, since this parameter was scaled to a mean of zero and standard deviation of one for all LID levels. For the *a*-parameter, the summary statistics are very similar with the exception of those for the high LID level for which the mean and standard deviation are both somewhat higher. This is consistent with the findings of Masters (1988) and Yen (1993) who observed that for positive LID, the relationship between some items is strengthened, thereby strengthening the relationship between the item and the total test. This results in an inflation of the *a*-values. The correlation of the *a*-parameter with the true parameters reported in Table 14 is quite high for the zero through medium LID levels (.91 to .88), but drops to a very low value of .256 for the high LID level.

TABLE 13
Summary Statistics for Item Parameter Estimates

	Level of LID				
	True	Zero	Low	Medium	High
a-parameter					
Mean	0.6585	0.6583	0.6627	0.6597	1.0031
S.D.	0.1682	0.1675	0.1756	0.1410	0.3692
b-parameter					
Mean	0.3275	0.3925	0.3678	0.2843	0.0742
S.D.	1.0611	1.0805	1.0572	1.0318	0.7236
c-parameter					
Mean	0.2005	0.2189	0.2135	0.1884	0.1082
S.D.	0.0765	0.0667	0.0706	0.0627	0.0689

TABLE 14
Pearson Correlation Coefficients among Item Parameters

		a- parameter		
LID Level	Zero	Low	Medium	High
True	0.9106	0.9084	0.8815	0.2563
Zero		0.8060	0.8218	0.2423
Low			0.8126	0.1939
Medium				0.2651

		b-parameter		
LID Level	Zero	Low	Medium	High
True	0.9934	0.9889	0.9888	0.9445
Zero		0.9878	0.9874	0.9509
Low			0.9841	0.9484
Medium				0.9484

		c-parameter		
LID Level	Zero	Low	Medium	High
True	0.8756	0.7978	0.7960	0.2263
Zero		0.7620	0.7722	0.2569
Low			0.7069	0.3251
Medium				0.3872

For the *c*-parameters, the standard deviations remain fairly constant while the mean values decrease. For the *b*-parameters, the means and standard deviations decrease as the LID increases. The correlation of the estimated *b*-parameters with the true *b*-parameters is not as alarming as that observed for the *a*-parameters, with this correlation dropping to only .94 for the high LID level. The correlations of the estimated *c*-parameters with the true *c*-parameters are lower overall than those observed for the *a*- and *b*-parameters, but such instability is typical for this parameter. This correlation does, however, drop sharply to .226 for the high LID level.

The correlation coefficients in Table 15 indicate that the relationship with the true ability parameters decreased slowly as the LID level increased, but dropped sharply for the high LID level.

TABLE 15
Pearson Correlation Coefficients among Ability Values

	Level of LID			
	Zero	Low	Medium	High
True	0.9573	0.9445	0.9115	0.6396
Zero		0.9238	0.8877	0.6352
Low			0.9008	0.6769
Medium				0.8371

Analysis of the Item and Test Characteristic Curves

Up to this point, findings have been discussed with respect to the effect of LID on the item and ability parameters. However, it is possible that looking at the *a*-, *b*-, and *c*-parameters separately may be misleading as these parameters co-vary in practice. The *a*-, *b*-, and *c*-parameters work together in equation 1 to produce an item characteristic curve (ICC) for an item. Therefore, studying the effect of LID on, say, the *a*-values is not very meaningful without considering the corresponding changes in the *b*- and *c*-parameters for a particular item. Various combinations of *a*-, *b*-, and *c*-parameters can produce ICCs that appear to be similar throughout most of the range of ability. Therefore, in this section, the impact of the various simulated degrees of LID on the item and test characteristic curves will be described and discussed.

Figures 2 through 4 present three sets of examples of overlay plots of the ICCs for each of the four LID levels. These plots tie together the results cited so far. Figure 2 presents the overlay plots for Item 5, which displayed only mild effects as a result of the increase in LID. For the zero and low LID levels, the ICCs are very similar. For the medium LID level, a slight crossing of the ICCs begins to emerge, with the estimated ICC dipping slightly below the true ICC at the lower end of the ability scale and rising slightly above the true ICC at the high end of the ability scale. At the high LID level, the ICCs diverge slightly more and in the same direction as for the medium LID condition.

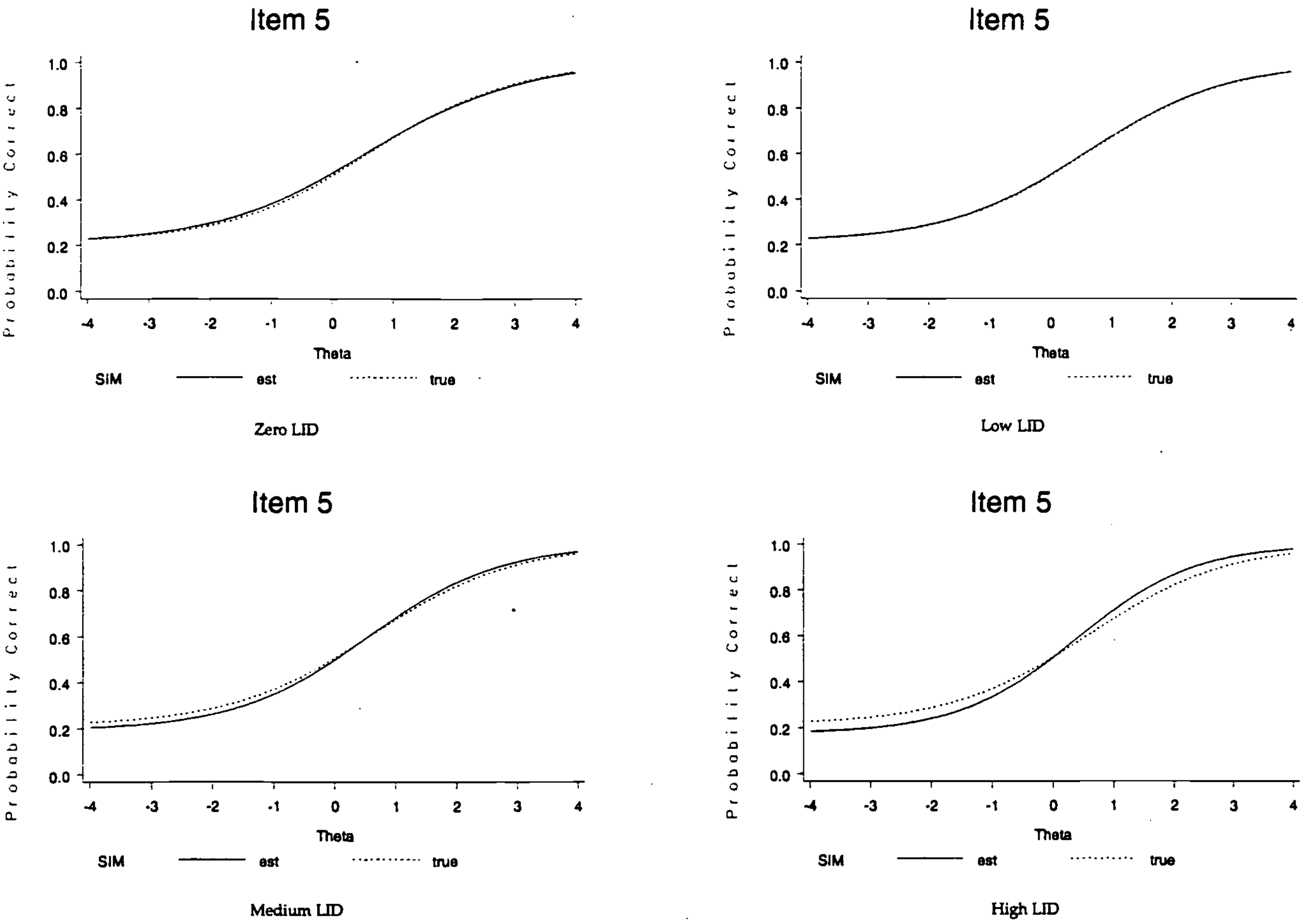


FIGURE 2. *Item characteristic curve overlay plots for item 5 that exhibited a mild effect*

Figure 3 represents the typical effect displayed for the ICCs at the various LID levels. Here, slight effects are observed even at the low LID level, but the effect becomes stronger as the LID is increased. Finally, Figure 4 shows a strong effect of the LID, particularly for the high LID level. A large number of items showed this degree of effect for the high LID. These effects highlight the underestimation of the *c*-parameters, as the success probabilities for low scoring test takers are underestimated. The overestimation of the *a*-parameters is also evidenced by the increase in the steepness of the ICC.

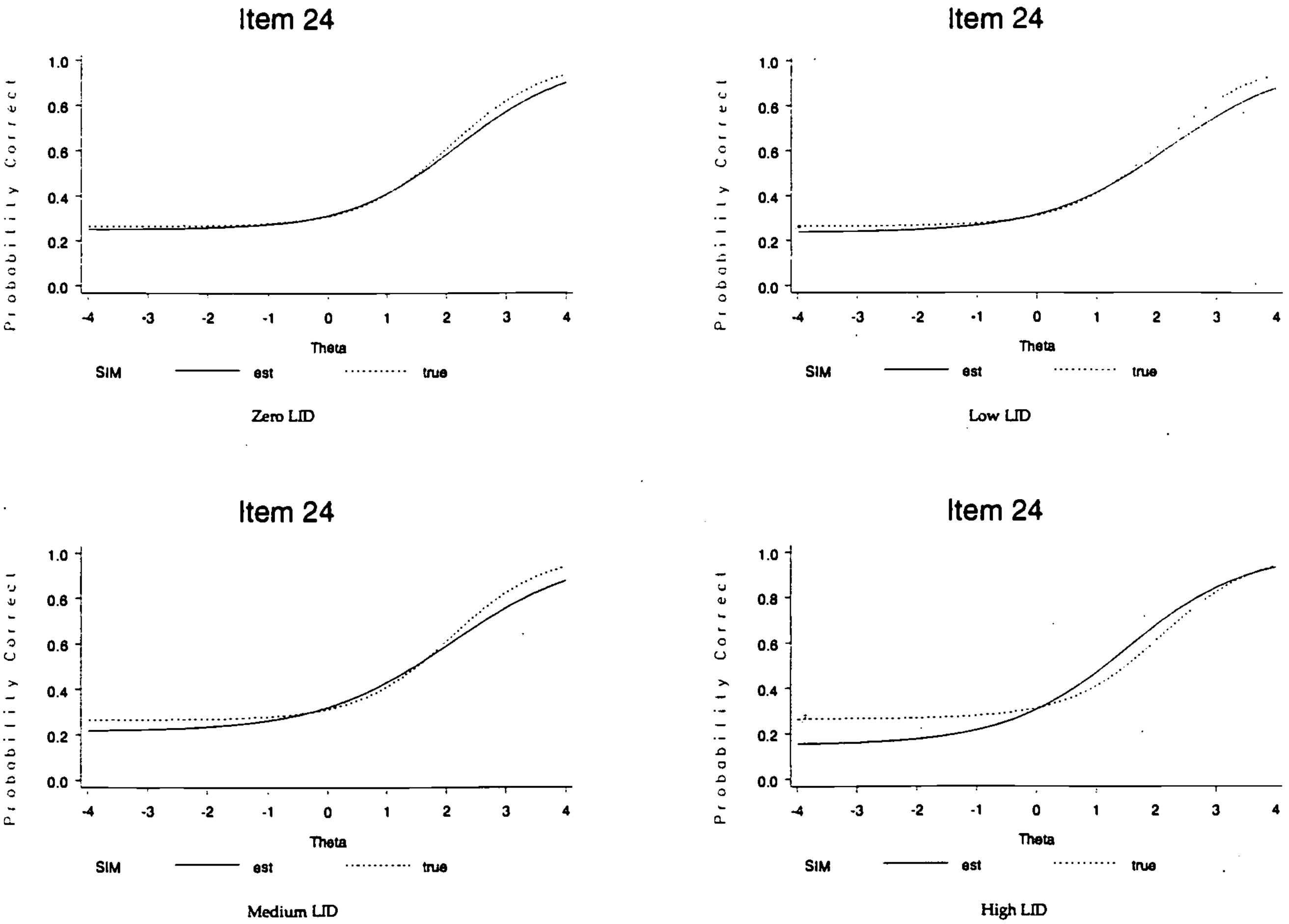
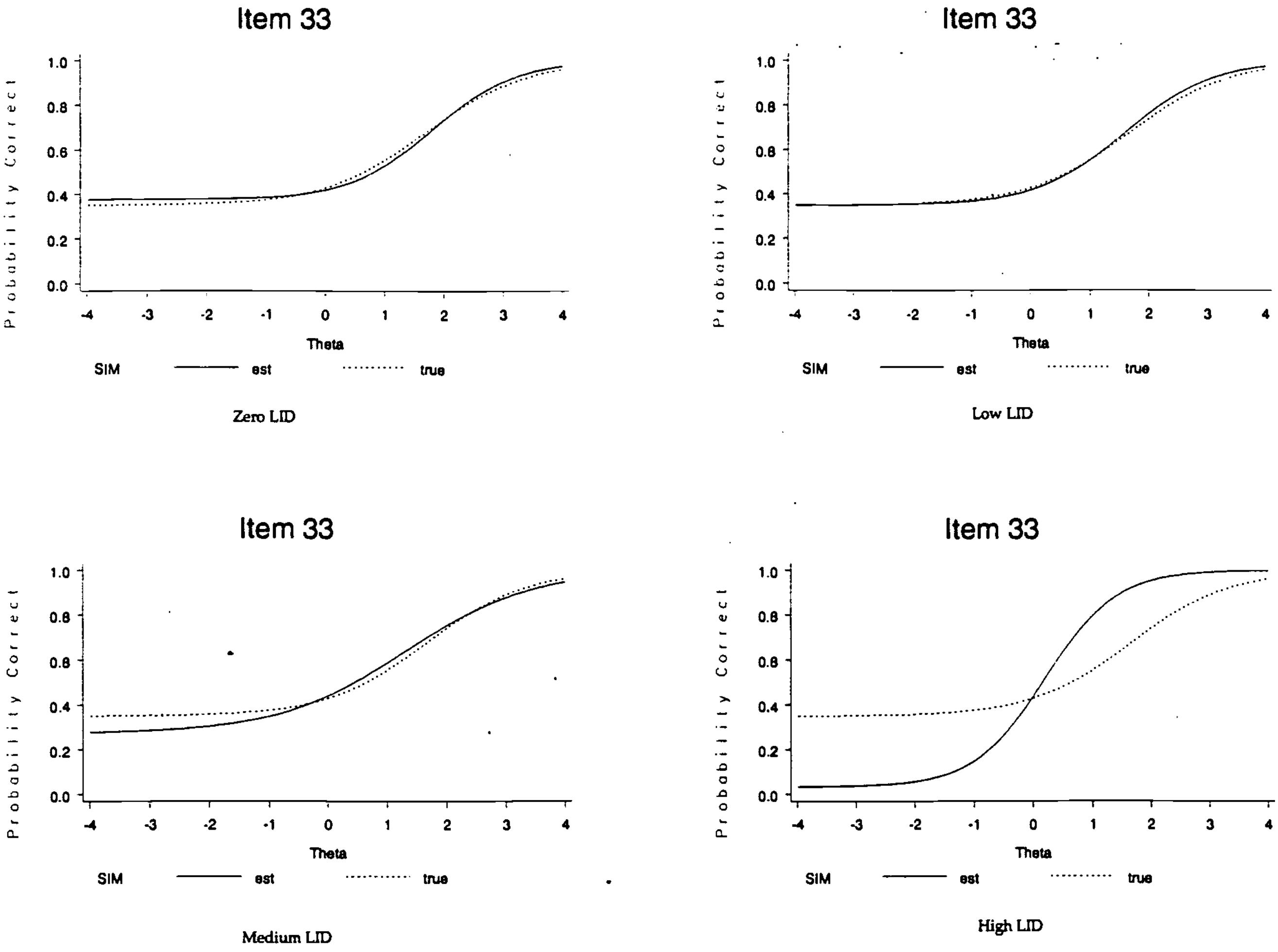


FIGURE 3. *Item characteristic curve overlay plots for item 24 that exhibited an average effect*



Zero LID

Low LID

Medium LID

High LID

FIGURE 4. *Item characteristic curve overlay plots for item 33 that exhibited a strong effect*

The test characteristic curve (TCC) overlay plot presented in Figure 5 demonstrates the effects of the LID over the entire test. For the zero, low, and medium LID levels, there is essentially no difference from the true TCC. For the high LID level, however, the underestimation of low scores and the overestimation of high scores is readily apparent.

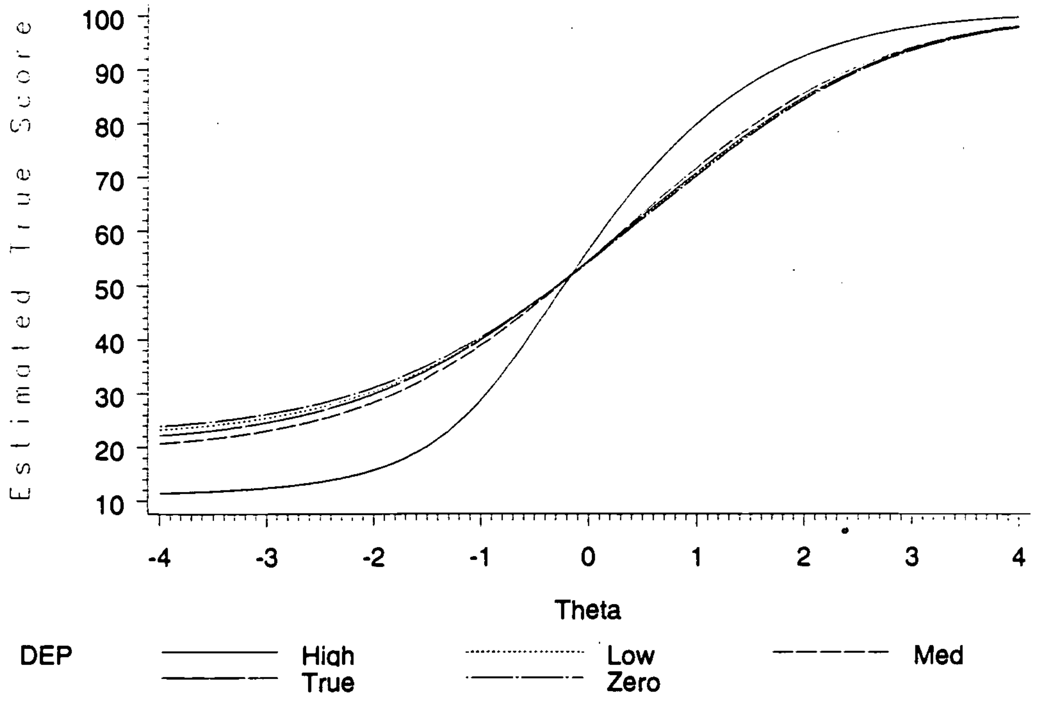


FIGURE 5. *Test characteristic curve overlay plot*

The Effect of Local Item Dependence on Scores

Up to this point, the effect of LID was analyzed at the item level. However, it was also important to determine the effect of LID on test scores. In order to study the effect of LID at this level, both score distributions and rank order correlations were evaluated.

Score Distributions

While studying the effects of LID on item parameters and ICCs are interesting and useful, many practitioners are also concerned with the effect of LID on score distributions. The overlay plots of the score distributions for the four levels of LID presented in Figures 6 through 9 provide global pictures of the effect of LID on this outcome measure. In each of these figures the true score distribution derived using the true item and ability parameters, the estimated true score distribution derived using the item and ability parameters estimated for the simulated data, and the observed score distributions are overlayed on a single plot. The estimated true and true score distributions were derived by applying Lord's (1980) method for estimating score distributions. The observed score distribution was derived by calculating a number-right score for each test taker and then calculating the frequency distributions for these scores.

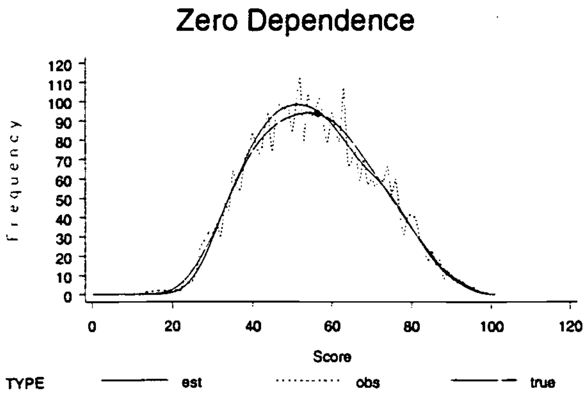


FIGURE 6. *Frequency distribution overlay plot for the zero LID level*

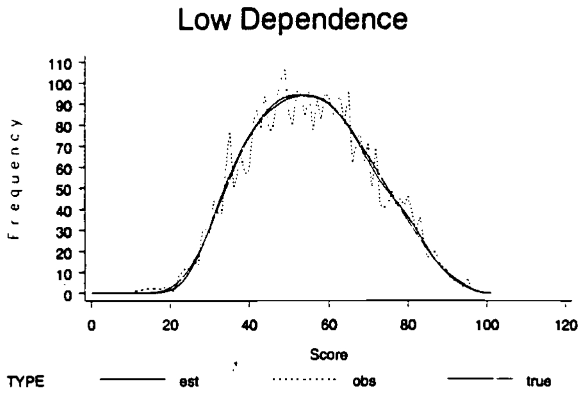


FIGURE 7. *Frequency distribution overlay plot for the low LID level*

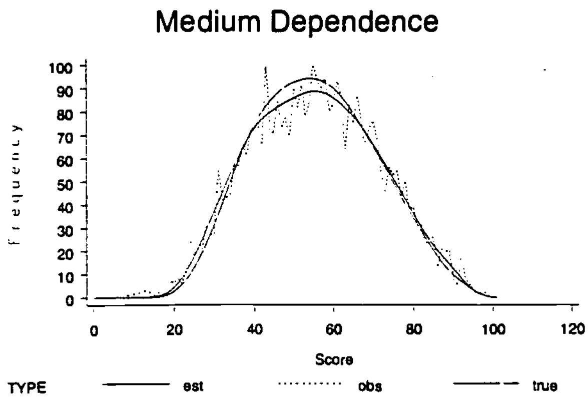


FIGURE 8. *Frequency distribution overlay plot for the medium LID level*

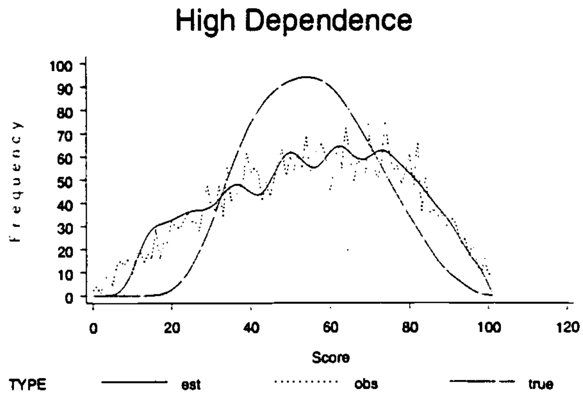


FIGURE 9. *frequency distribution overlay plot for the high LID level*

For the zero LID level presented in Figure 6, the three distributions are very similar. The estimated true score distribution is shifted slightly to the right and peaks slightly higher as compared to the true distribution, but this difference appears to be nonsignificant. For the low LID level in Figure 7, the three distributions are practically identical. Figure 8 demonstrates that for the medium LID level, some slight differences begin to emerge, with the observed and estimated true score distributions peaking somewhat lower than the true score distribution. Figure 9 shows that at the high LID level, the differences become dramatic. The observed and estimated true score distributions are still very similar to one another, but they lose their normality and do not resemble the true score distribution. The distributions for the high LID data spread out for reasons related to what was observed for the item characteristic curves in the previous section. Again, the scores of low ability test takers are underestimated and the scores of high ability test takers are overestimated. This effect causes the score distribution to spread out at the tails and flatten in the middle.

Rank Order

Table 16 presents the Spearman rank order correlation coefficients among the true, estimated true, and observed scores for the various levels of LID. These correlations are all .9 and over until the high LID level is reached. Here, the correlations between the rank orderings drop to .625 for the true and estimated true scores and .638 for the true and observed scores. The rank order correlation between the observed and estimated true scores remains high at .973. This indicates that when the LID becomes severe, not only are high scores overestimated and low scores underestimated, but the relative standing of individuals is also affected.

TABLE 16
Spearman Rank Order Correlations

Level of LID		True	Observed
Zero	Estimated True	0.9598	0.9942
	True		0.9547
Low	Estimated True	0.9448	0.9943
	True		0.9374
Medium	Estimated True	0.9115	0.9953
	True		0.8956
High	Estimated True	0.6254	0.9730
	True		0.6383

Conclusions and Future Directions

The impact of LID was explored for both the calibration results and score distributions. Some observations regarding the Q_3 statistic were also made. This section discusses some conclusions that may be drawn based on the findings reported here. The implication of these findings for measurement are also discussed, and some future directions are suggested.

Impact of Local Item Dependence on Calibration Results

Overall, the results observed for the calibration of the dependent data revealed that violations of the local item independence assumption cause low scores to be underestimated and high scores to be overestimated. This effect has the expected impact on the IRT calibration results. The underestimation of low scores causes the *c*-parameter to drop, and the increased steepness of the item characteristic curve causes the *a*-parameter to become inflated. The *b*-parameter tends to be underestimated. The item and test characteristic curve overlay plots demonstrate these outcomes most effectively. The impact was mainly observed for the high LID level, while the effects for the low and medium LID levels appeared to be minimal. Fortunately, the LSAT appears to exhibit at most only medium levels of LID.

At the medium LID level, some mild effects of the LID were observed. This would imply that the effect of violations of local item independence on LSAT calibration results should be monitored, but the effects are not likely to be problematic. The effects of high LID on the calibration results are very problematic, and the application of unidimensional IRT to data displaying this level of LID may not always be appropriate. Again, it should be noted that the Q_3 statistic has a tendency to underestimate high LID. Therefore, it is possible that the LSAT does have a higher level of LID than was defined here as "medium," but the LID was underestimated by the Q_3 statistic.

Impact of Local Item Dependence on the Score Distribution

The effects of LID observed for the calibration results manifest themselves in predictable ways in the score distributions. As LID levels were increased, the underestimation of low scores and overestimation of high scores caused the score distribution to be spread out at the tails and flatten in the center. This effect was observed most clearly for the high LID level. One very encouraging result was observed, however, with respect to the score distributions. While a high level of LID caused the observed and estimated score distributions to depart dramatically from the true distribution, the observed score distribution was modeled quite well by the score distribution predicted from the dependent IRT parameters. Therefore, when our purpose is to model observed data, IRT performs quite well, even in cases of extreme LID.

Again, these results imply that for tests displaying only a low degree of LID, there is no reason for concern. Even for the medium LID level, representing the LSAT, the results do not appear to be problematic. It is at the high LID level that these effects become troublesome for certain purposes.

The Q3 Statistic

The results presented in this study suggest that the Q_3 statistic should be studied further. The simulation portion of this study indicated that for medium and high LID levels, the true LID was not adequately reflected by the Q_3 statistic after the data were calibrated. This result is due to the fact that the item and ability parameter estimates, which are contaminated by the LID of the data, are used in the calculation of the statistic. It seems logical that as the LID level increases, this contamination becomes more extreme, and the true LID is underestimated. This problem is greater when a high level of LID exists for a large block of items. As was discussed in the results section, perhaps the correction factor of $-1/(n-1)$ suggested by Yen is appropriate only in cases of zero and low LID and is inappropriate when the dependence becomes greater. This also implies that the results observed for the real-data analyses may represent an underestimate of the true state of LID. Throughout this study, the high LID level has been described as an extreme case of LID that would only be observed in rare situations. However, it is possible that the higher levels of LID that do exist in the real data have been underestimated and the high LID level simulated is not as rare as implied here. Further research should be carried out in order to investigate this problem.

Implications for Measurement

The results observed and discussed here indicate that applying unidimensional item response theory to test data that exhibit a high degree of LID may not always be appropriate. For any test displaying the highest level of LID studied here, the application of IRT would unfairly advantage some test takers and disadvantage others. Even at the medium LID level, some degree of caution should be applied. While the effects at this level are not as dramatic as those observed for the high LID level, some effect is readily apparent. In any case, the degree of violation of the local item independence assumption should be investigated for any operational testing program, as there are likely to be many tests that display a degree of LID that falls somewhere in between the medium and high LID levels studied here. As has been discussed, there are also likely to be some testing situations in which the level of LID is in fact as high as the high LID level. For these cases, the application of IRT is obviously problematic.

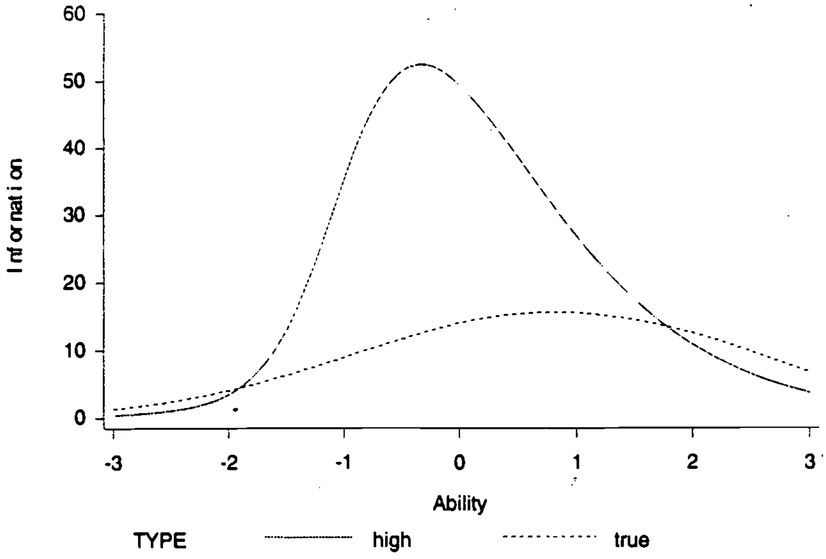


FIGURE 10. *Test information function overlay plot*

In terms of various IRT applications, these results are very problematic. The LSAT is assembled to match target test information functions and target score distributions. While information functions were not studied here, it has been shown that LID causes an overestimation of this measure (Yen, 1993). Figure 10 overlays the test information functions for the true parameters and the parameters derived for the high LID level. The overestimation of this measure for high LID is clearly apparent. A test assembled to match target test information curves when the local item independence assumption is being violated to a significant degree would result in a test that is not providing as much information as is intended.

In terms of a target score distribution, an assembly based on essentially independent data might result in a test with higher LID due to context effects. The test thus assembled might not have the intended IRT characteristics due to LID influences.

Possibly the widest current application of IRT is to test equating, and the effect of LID has clear implications for this procedure. LSAC equates the LSAT using IRT true-score equating. Consider, for example, the case where an operational test form is administered to a group of test takers along with a test section containing pretest items. A situation may arise where the operational form has a high level of LID while the pretest section does not. When the operational form and the pretest section are calibrated together, the operational items will have the greatest impact on the ability scale. Thus, the LID will have a contaminating effect upon the experimental items even though they do not show a high degree of LID on their own. In this particular situation, the common population has a contaminating effect on the pretest items even though the LID among these items is not problematic.

In the application of IRT to computerized adaptive testing (CAT), violations of the local item independence assumption becomes a serious issue. Here, the test taker is presented with test items via computer administration. As the test progresses, information functions are derived for the available items, and the item providing the maximum information at the current estimate of the test taker's ability is selected for administration. Usually, when the standard error of the test taker's ability estimate has been lowered to a predefined level, the testing session is terminated. Since the standard error in IRT is the reciprocal of information, the overestimation of information discussed earlier (see Figure 10) is clearly a problem. While the information function is overestimated, the standard error is underestimated, and the test taker's ability is not being estimated with as much precision as we think. Perhaps the greatest problem with CAT is that the effect of LID is difficult to assess since each test taker responds to a different set of items that may in combination yield varying degrees of LID. Some researchers have been addressing this problem with the use of testlets or item bundles. The potential for inequities in the CAT environment should definitely be addressed in some way.

Future Directions

The results observed here make it clear that much more research is needed on this vital assumption of IRT. More research should be carried out with the Q_3 statistic in order to determine how to interpret this statistic more clearly or to improve upon it.

While this study thoroughly investigated the overall effects of violations of the local item independence assumption on the calibration results and score distribution, the impact of this problem for the individual test taker was not directly investigated. The results observed for the Spearman rank order correlation coefficient revealed that the relative standing of test takers is affected by LID. Future research should explore the extent to which the percentile ranks of test takers at different points along the ability scale are changed from their true percentile ranks when LID is introduced into the data.

As was mentioned above, the results observed have implications for the use of IRT in many testing applications. The effect of LID on test assembly based on both target information functions and target test characteristic curves should be explored. Also, the impact of LID on IRT equating should be investigated thoroughly. Finally, future research should proceed toward equalizing the effects of LID for test takers in the CAT environment.

References

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence* (RR-87-14). Iowa City, IA: ACT.

Ackerman, T., & Spray, J. (1987). *A general model for item dependency* (RR-87-9). Iowa City, IA: ACT.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-573.

Andrich, D. (1985). A latent trait model for items with response dependencies. In S. E. Embretson (Ed.), *Test design: Developments in psychology and psychometrics*. Orlando, FL.: Academic Press.

Bell, R. C., Pattison, P. E., & Withers, G. P. (1988). Conditional independence in a clustered item test. *Applied Psychological Measurement, 12,* 15-26.

Embretson (Whitely), S. (1984). A general latent trait model for response processes. *Psychometrika, 49,* 175-186.

Goldstein, H. (1980). Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology, 33,* 234-246.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer.

Holland, P. W. (1981). When are item response models consistent with observed data? *Psychometrika, 46,* 79-92.

Jannarone, R. J. (1986). Conjunctive item response theory kernels. *Psychometrika,* 51, 357-373.

Jannarone, R. J. (1987). *Locally dependent models for reflecting learning abilities* (Center For Machine Intelligence Report No. 87-67). Columbia, SC: University of South Carolina.

Jannarone, R. J. (1991a). Conjunctive measurement theory: Cognitive research prospects. In M. Wilson (Ed.), *Objective measurement: Theory into practice, Volume 1*. Norwood, NJ: Ablex Publishing Corporation.

Jannarone, R. J. (1991b). *Locally dependent cognitive process measurement, contrasts and connections with traditional test theory*. Unpublished manuscript, University of South Carolina.

Jannarone, R. J. (1991c). *Measuring quickness and correctness concurrently: A conjunctive IRT approach*. Unpublished manuscript, University of South Carolina.

Jannarone, R. J. (in press). Local dependence: Objectively measurable or objectionably abominable? In M. Wilson (Ed.), *Objective measurement: Theory into practice, Volume 2*. Norwood, NJ: Ablex Publishing Corporation.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika, 49,* 223-245.

Kempf, W. F. (1977). A dynamic test model and its use in the microevaluation of instructional material. In H. Spada & W. F. Kempf (Eds.), *Structural models of thinking & learning*. Vienna: Hans.

Lord, F. M. (1953). The relation of test score to the trait underlying the test. *Educational and Psychological Measurement, 13,* 517-549.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems. Hillsdale,* NJ: Erlbaum.

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.

Pashley, P. J., & Reese, L. M. (1995). *On generating locally dependent item responses*. Unpublished manuscript.

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

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika, 53,* 349-359.

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

Spray, J. A., & Ackerman, T. (1987). *The effect of item response dependency on trait or ability dimensionality* (RR-87-10). Iowa City, IA: ACT.

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

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

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

Wollenberg, A. L. van den. (1982). Two new test statistics for the Rasch model. *Psychometrika, 47,* 123-140.

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*(2), 125-145.

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

NOTICE

Reproduction Basis