DOCUMENT RESUME

ED 469 173 TM 034 478

AUTHOR	Reese, Lynda M.; Pashley, Peter J.
TITLE	Impact of Local Item Dependence on True-Score Equating. LSAC Research Report Series.
INSTITUTION	Law School Admission Council, Newtown, PA.
REPORT NO	LSAC-RR-97-01
PUB DATE	1999-08-00
NOTE	13p.
PUB TYPE	Numerical/Quantitative Data (110) -- Reports - Research (143)
EDRS PRICE	EDRS Price MF01/PC01 Plus Postage.
DESCRIPTORS	College Entrance Examinations; *Equated Scores; *Item Response Theory; Law Schools; *Test Items; *True Scores
IDENTIFIERS	*Item Dependence; *Law School Admission Test; Local Independence (Tests)

ABSTRACT

This study investigated the practical effects of local item dependence (LID) on item response theory (IRT) true-score equating. A scenario was defined that emulated the Law School Admission Test (LSAT) preequating model, and data were generated to assess the impact of different degrees of LID on final equating outcomes. An extreme amount of LID was induced for a dataset that was analyzed by applying the Rasch model, and a typical amount of LID was induced for a dataset that was analyzed by applying the three-parameter logistic model. Results indicate that the effects of the extreme LID on the conversion line derived after scaling was carried out were negligible. Effects on the observed score distribution were noted for the extreme LID case, however, and such effects would have implications for score reporting. For the typical amount of LID, both the equating conversion line and the observed score distribution were essentially unaffected. (Author/SLD)

TM

Impact of Local Item Dependence on True-Score Equating

Lynda M. Reese and Peter J. Pashley
Law School Admission Council

Law School Admission Council
Research Report 97-01
August 1999

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 assembly, evaluation, and equating of new LSAT forms, item response theory (IRT) is applied. IRT is a mathematical model that relates the probability that a test taker will answer a single test item (i.e., 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 test takers 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, if a test section is especially difficult, fatigue may adversely affect the performance of test takers on the items at the end of the section. In this case, the difficulty level of the items found at the beginning of the section affect performance on later items, and so these items are said to exhibit some degree of local item dependence (LID).

An IRT equating method called true-score equating is routinely applied for the LSAT. Equating is a process that adjusts for minor differences in difficulty between test forms to assure that the score reported for a test taker carries the same meaning regardless of the particular form of the test administered to that test taker. Previous research by Pashley and Reese and Reese has demonstrated that an extreme amount of LID has an impact on certain IRT outcomes that are related to this equating process. Since test equating is so important to the assurance that the LSAT is equitable for all test takers, the study described herein explored the impact of LID on the IRT true-score equating process routinely carried out for the test.

The results indicated that even for an extreme amount of LID, the equating process eliminated the effect of the LID on the final conversion table. However, the extreme level of LID studied here did have an impact on the distribution of number-correct scores. Since number-correct scores are used in score reporting for the LSAT, this result could be a concern if the LSAT were to display such an extreme amount of LID. Fortunately, for the amount of LID typically displayed for the LSAT, no adverse impact was observed for either the conversion table or the distribution of number-correct scores.

Abstract

This study investigated the practical effects of local item dependence (LID) on item response theory (IRT) true-score equating. A scenario was defined that emulated the LSAT section preequating model, and data were generated to assess the impact of different degrees of LID on the final equating outcomes. An extreme amount of LID was induced for a dataset that was analyzed by applying the Rasch model, and a typical amount of LID was induced for a dataset that was analyzed by applying the three-parameter logistic model. Results indicated that the effects of the extreme LID on the conversion line derived after scaling was carried out were negligible. Effects on the observed score distribution were noted for the extreme LID case, however, and such effects would have implications for score reporting. For the typical amount of LID, both the equating conversion line and the observed score distribution were essentially unaffected.

Introduction

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

$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}}, \quad (1)$$

where , a_i, b_i, and c_i represent item i's discrimination, difficulty, and pseudo-guessing parameters, respectively, θ represents the ability level of the test taker, and D is a scaling factor usually set to 1.7 (Lord, 1980). 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

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

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.

Previous research by Pashley and Reese (1999) and Reese (1995) explored the impact of local item dependence (LID) on some commonly applied IRT outcomes. Until this work, most researchers investigating the effects of LID did so by analyzing multidimensional data. Since the unidimensionality and local independence assumptions of IRT are related, data that are multidimensional will exhibit LID. However, the exact structure of the LID for a multidimensional dataset is not interpretable. The research of Pashley and Reese (1999) and Reese (1995) represents a departure from this common practice in that the LID data were generated directly from correlation matrices that defined the structure of the LID for the dataset. This research revealed that item and test characteristic curves and score distributions were strongly affected when an extreme level of LID was induced in the data.

During the time period the above research was being conducted, De Champlain (1995) compared LSAT IRT true-score equating results for the Caucasian, African American, and Hispanic subgroups and found that their separate conversion lines were statistically equivalent. These three conversion lines were also statistically equivalent to the conversion line derived for the total group. These results were observed despite the fact that dimensionality analyses suggest a two-dimensional model for the Caucasian and African American subgroups, while a more complex model is needed for the Hispanic subgroup.

The Pashley and Reese, Reese, and De Champlain studies served as the impetus for the current study. The De Champlain results imply that the dimensionality structure of a dataset does not have an impact on the IRT true-score equating results. In light of the relationship between the unidimensionality and local item independence assumptions of IRT, this study was carried out to investigate the practical effects of LID on IRT true-score equating results.

Methodology

Study Design

This study was designed to represent a simplified version of the current LSAT section pre-equating model, and a degree of LID was introduced into the data at a certain point in the model. As described in Table 1, data were generated to represent the administration of a 50-item operational form (Form I) and a 50-item preoperational form (Form II). At this first administration, both Form I and Form II were defined to have zero LID. Next, data were generated to represent a later administration at which Form II was administered operationally while Form I was administered preoperationally. For the second administration, a degree of LID was defined for the Form I and II items. For both administrations, the operational form was administered before the preoperational form. Finally, preoperational Form I was scaled to operational Form I by applying the characteristic curve scaling method of Stocking and Lord (1983), treating the 50 Form II items as common items.

TABLE 1
Description of Study Design

OP I 50 items	PO II 50 items
OP II′ (before scaling) 50 items	PO I′ (before scaling) 50 items
OP II (after scaling) 50 items	PO I (after scaling) 50 items

While the circular nature of this study design is not completely realistic from an operational point of view, this design allowed the factors of interest in this study to be manipulated. In this design, Forms I and II displayed zero LID at one administration and a degree of LID at a subsequent administration. Such a situation could arise in practice. For instance, if Form II was slightly more difficult than Form I, this difference in difficulty may not have had an effect when Form II was administered after Form I. However, LID could result from fatigue when the more difficult form was presented first.

Data Generation

To create simulated data, the method described by Pashley and Reese (1999) was applied. Specifically, the steps carried in the data generation were as follows:

Step 1: A desired correlation structure among the items was defined.

Step 2: A vector x of multivariate random deviates was generated according to the correlation structure defined in step 1.

Step 3: Using the inverse normal transformation, x was transformed to y, a vector of uniform (0,1) deviates.

Step 4: Each uniform deviate was compared to individual item correct probabilities in order to obtain 0 or 1 item responses.

Rasch Model Analyses

In the first set of analyses carried out, the most simplistic IRT model, called the Rasch model, was used. Here, an item is described only by its difficulty, or b-parameter value. For this model, the equation used to define the probability that a test taker of a specific ability level will answer an item correctly is defined by the equation

$$P_i(\theta) = \frac{1}{1 + e^{-D\bar{a}_i(\theta - b_i)}}, \tag{3}$$

where all terms are defined as for Equation 1, and $\bar{a}$ represents a common value for the a-parameter.

In this set of analyses, a sample of 100 standard normal b-parameter values was generated and treated as the true b-parameter values. A common value of 1.0 was used for the a-parameter. Responses to these items were generated for 1,000 standard normal ability values as described in the data generation section above. For the first administration, zero LID was induced for all items. For the second administration, an extreme level of LID was induced for Forms I and II. This extreme level of LID was comparable to the high LID level of the Reese (1995) study and was defined to represent the LID of a highly dependent item set. Such an extreme level was induced so that we could assure that an effect would be observed.

Three-parameter Model Analyses

In order to evaluate a set of conditions that would more realistically represent the LSAT, analyses were also carried out applying the IRT three-parameter logistic model. This is the IRT model currently used in equating the LSAT, and the probability that a test taker of a specific ability level will answer an item correctly is defined by Equation 1.

Item parameter values (*a*-, *b*-, and *c*-parameters) for a sample of 100 typical LSAT items were treated as the true item parameter values. Responses to these items were generated for 1,000 standard normal ability values as described in the data generation section above. For the first administration, zero LID was induced for all items. For the second administration, a typical level of LID was induced for Forms I and II. This level of LID was comparable to the medium LID level of the Reese (1995) study and was defined to represent what is typically observed for the LSAT. This example should provide a realistic picture of the impact of LID on the LSAT true-score equating.

Data Calibration

The simulated data for both IRT models were calibrated using BILOG (386 BILOG 3, Mislevy & Bock, 1990). The default scoring method, number of quadrature points, and priors were utilized. To assure that the item and ability parameter estimates for the true and generated data were on a common scale, the *rescale* option in BILOG was applied, scaling the ability parameters from each calibration to have a mean of zero and a standard deviation of one.

Results

Rasch Model Analyses

Figure 1 displays the equating results for preoperational Form I for the Rasch model. Recall that an extreme level of LID was induced for Forms I and II in this example. This figure reveals that before the scaling was carried out, the scores of the lower scoring test takers were underestimated while the scores of the higher scoring test takers were overestimated as a result of this level of LID. However, after scaling was carried out, the effects of the LID nearly disappeared, and the extreme scaled conversion line came very close to those for the true item parameters and the zero LID data.

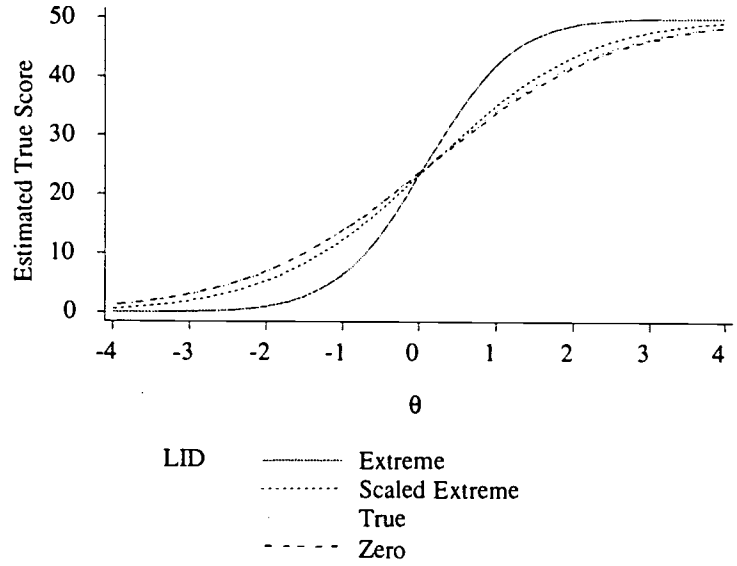

FIGURE 1. *Test characteristic curves for the Rasch model analyses*

Figure 2 overlays the frequency distribution of the number-correct score for each of the simulated administrations of Form I. Note that the distribution of these scores for operational Form I is normal in shape, while the scores for preoperational Form I are more spread out. This figure demonstrates that the LID induced in the second administration had an effect on the number-correct scores of the Form I simulated test takers, causing the scores to be spread toward the extremes of the score scale.

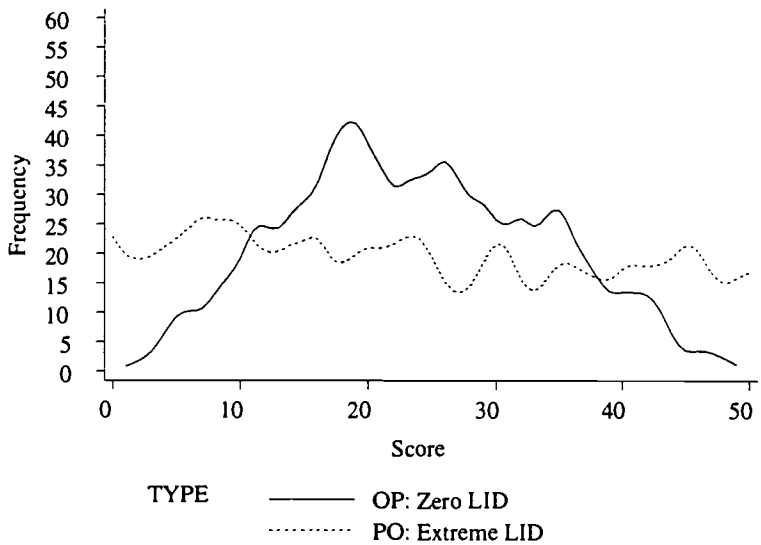

FIGURE 2. *Observed score distributions for the Rasch model analyses*

Three-parameter Model Analyses

Figure 3 displays the equating results for the three-parameter model analyses. Recall that for this case, a typical amount of LID was induced for the second administration. Here, the Form I conversion lines before and after the scaling are practically identical and are practically identical to the true and zero LID parameters. This indicates that for a typical amount of LID, there is no adverse effect on the equating conversion line. The observed score distribution for the typical LID three-parameter case, displayed in Figure 4, is equally encouraging. Here we see that the observed score distributions for both Form I administrations are essentially identical.

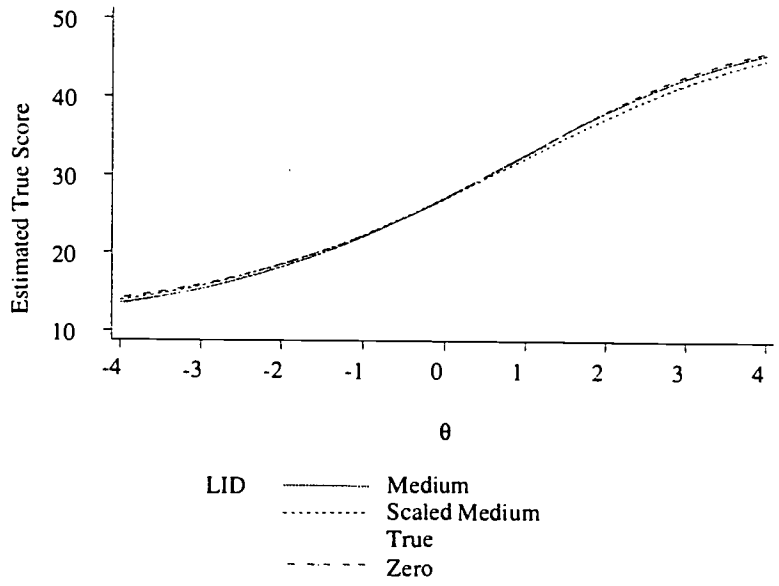

FIGURE 3. *Test characteristic curves for the three-parameter model analyses*

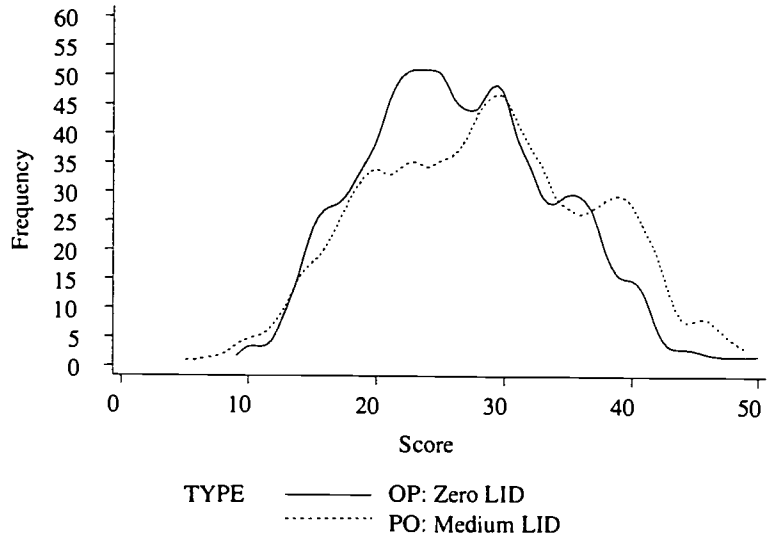

FIGURE 4. *Observed score distributions for the three-parameter model analyses*

Discussion

As stated in the introduction section, the results observed in the Pashley and Reese (1999) and Reese (1995) studies served as the major impetus for the current study. Here, four levels of LID were defined, and associated datasets were generated by applying the data generation routine described above and discussed more fully in Pashley and Reese (1999). Four LID levels were explored in this research, denoted as zero, low, medium, and high. Test characteristic curves for each of these LID levels are presented in Figure 5. This figure demonstrates that the high level of LID caused lower scores to be under estimated and higher scores to be over estimated. Since the test characteristic curve is the final product of IRT true-score equating, this result raised a concern over whether the process of IRT true-score equating could correct for this effect in a typical section preequating design.

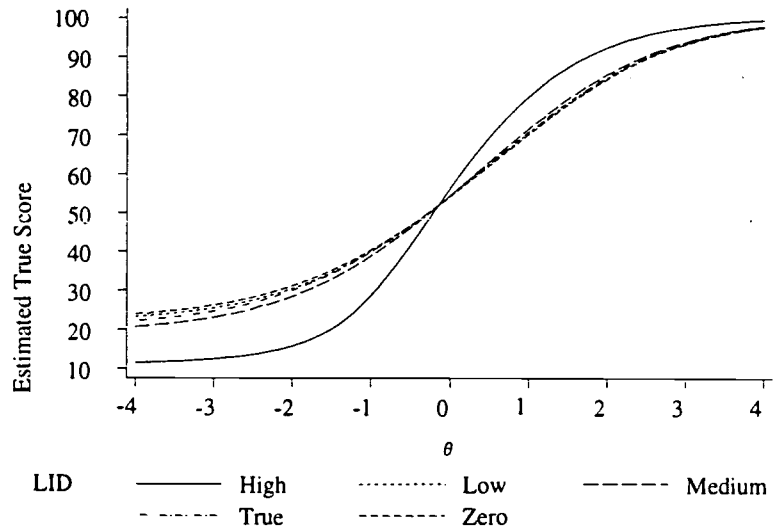

FIGURE 5. *Test characteristic curves for the zero, low, medium, and high LID levels*

The results observed here are very encouraging with regard to the effects of LID on the conversion line produced through IRT true-score equating. The level of LID defined for the Rasch model analyses was similar to the high LID level presented in Figure 5. Figure 1 demonstrates that before scaling is carried out, the test characteristic curve produced by this data mimics that observed in previous research. However, any effects on the conversion line nearly disappeared after scaling was carried out. Therefore, while a high level of LID has an extreme impact on the test characteristic curve, these effects are corrected through the scaling and equating process.

For the typical LID level simulated for the three-parameter model analyses, no effects on the conversion line were observed even before the scaling was carried out. Recall that this level corresponds to the medium LID level displayed in Figure 5. These results were especially encouraging since the LID level induced here was defined to represent what is typically observed for the LSAT.

The Reese (1995) study also explored the impact of LID on the score distribution, and the results observed for the medium and high LID levels are displayed in Figures 6 and 7, respectively. 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 for the simulated data, and the observed score distribution are overlayed on a single plot. The true and estimated true-score distributions were derived by applying Lord's (1980) method for predicting score distributions. The observed score distributions were derived by calculating the number-right score for each test taker and then calculating the frequency distribution for these scores.

Figure 6 shows that for the medium LID level, no adverse effects on the score distributions were observed. In contrast, Figure 7 shows an extreme effect, with the observed and estimated true-score distributions flattening in the center and spreading out to the extremes.

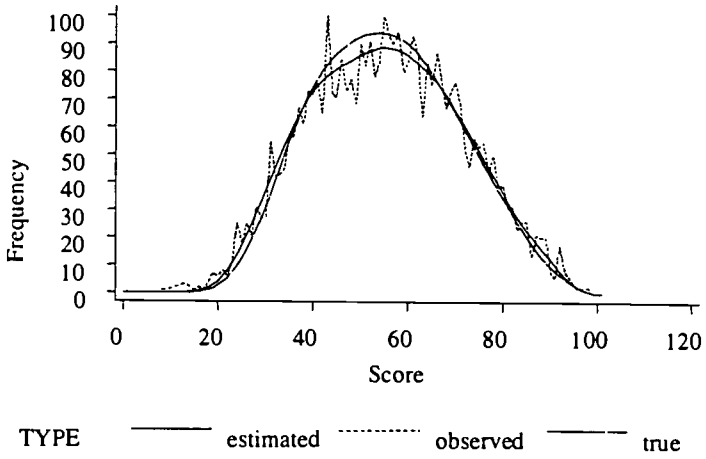

FIGURE 6. *Frequency distribution overlay plot for the medium LID level*

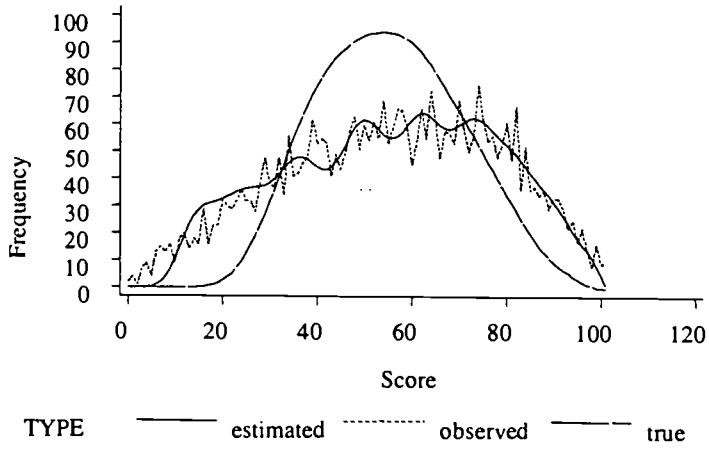

FIGURE 7. *Frequency distribution overlay plot for the high LID level*

The results observed for score distribution for the extreme LID level of this study were similar to those observed by Reese (1995). This result is troublesome since in creating a conversion table for score-reporting, it is a common practice to use the number-correct score as a surrogate for the estimated true-score, as this is more easily understood by test takers. In fact, this is the score reporting method employed for the LSAT. The results observed here indicate that such practices would not be equitable for all test takers if an extreme amount of LID were present in the data. Fortunately, the more typical level of LID induced for the three-parameter model analyses did not result in these adverse effects on the observed score distributions. Again, this is similar to what was observed by Reese (1995), as demonstrated in Figure 7. These results indicate that for the LSAT, the LID typically displayed for the test should not have any adverse impact on the current test equating practices.

References

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

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

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

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

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

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").