LSAC RESEARCH REPORT SERIES

On the Use of Collateral Item Response Information to Improve Pretest Item Calibration

William Stout, Terry Ackerman, Dan Bolt,
Amy Goodwin Froelich, Dan Heck
University of Illinois at Urbana-Champaign

Law School Admission Council
Computerized Testing Report 98-13
September 2003

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 201 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 these reports are those of the authors and do not necessarily reflect the position or policy of the Law School Admission Council.

Table of Contents

Executive Summary

An increasingly important issue in considering the feasibility of a computerized Law School Admission Test (LSAT) is the development of accurate statistical estimation tools that work well for the modest amounts of test taker data per test question ("item") that are typical of computerized tests. This is especially important for pretest items, whereby, currently, a pretest item is an item administered on an earlier exam as a nonoperational item in order to assess its characteristics so the item can be effectively used operationally on future LSAT administrations.

In computer adaptive testing, the rates at which items are exposed to test takers must be carefully controlled to ensure test security. Hence a very large operational item pool, whose items have already had their statistical properties analyzed at the pretest stage described above, must continuously be maintained and gradually renewed. Thus the goal of obtaining accurate estimates of characteristics (called "item parameter calibration") of pretest items by using as few test takers as possible for each pretest item becomes even more important in computer adaptive testing than in convential paper-and-pencil testing. In this study, several approaches for extracting useful collateral information for the purpose of pretest item parameter calibration are developed and studied. In the general statistical estimation context, collateral information refers to additional estimation information derived from variables that are distinct from, but correlated with, the studied relevant variable of interest.

That LSAT currently consists of three item types: logical reasoning (LR), analytical reasoning (AR), and reading comprehension (RC). Even though separate scores are not currently reported for the different item types, the reporting of separate scores for at least some item types could be considered for a computerized LSAT. Thus, for calibrating pretest items for one item type, the use of collateral information from some other item type could be helpful.

The project's specific goal is the investigation of the use of such collateral information to aid in reducing the number of test takers per pretest item required for calibrating test items. For the LSAT, collateral information may be of particular value in calibrating pretest items, because, as previous studies have suggested, the LSAT measures two dominant abilities ("dimensions"). One such dimension is defined by the AR items, and the second by the combined LR and RC items. While in practice, pretest RC items may be calibrated by considering only RC operational item responses, it is also possible to use test taker performances on AR items as well (i.e., collateral information from the AR items) to aid in the calibration of pretest RC items.

This study evaluates the practical benefit (if any) of using collateral information from one item type when statistically analyzing pretest items of some other item type. The criterion for evaluation of pretest item calibration accuracy was the reduction achieved by the use of collateral information in the number of test takers that must be administered each pretest item so that its item parameter estimates reach a specified level of accuracy. Our proposed methods involve both statistical tools that presume only one dominant ability being measured on a test and tools that assume two dominant abilities being measured, as has been confirmed to be true for the LSAT by statistical analyses as mentioned above. They also involve tools that include and exclude an intermediate ability estimation stage. These methods are described in detail and evaluated through simulation studies.

Results from simulation studies demonstrate that there is a practically important amount of improvement in calibrating pretest items of one item type attributable to the use of collateral information from some other item type provided only a moderate number of operational items are present for each item type. Unfortunately, for purposes of an LSAT computerized administration as currently contemplated by Law School Admission Council (LSAC) where the number of operational items per item type is expected to be more substantial, the high level of ability estimation accuracy for any one item type (because of the more substantial number of items being anticipated) seems to offset any advantage that could be gained through introducing collateral information from some other item type. However, if future LSAT pretest item calibration needs to evolve in any way where short substantively similar operational sets of items need to be calibrated, then our results suggest collateral information could provide practically important gains in pretest item calibration accuracy.

Introduction

The growing recognition that most standardized tests are multidimensional presents new statistical challenges in the application of item response models. The unidimensionality requirement of many IRT models implies that accurate parameter estimation is optimal only when the interaction between items and test takers can be well modeled by a single latent dimension. Thus tests known to be multidimensional may be better modeled by calibrating subsets of test items where each such subset is known to be relatively dimensionally homogeneous. This is not to imply that such calibrations must be performed independently,

however. In fact, when the latent dimensions are correlated, there is reason to believe that item responses from one such subset of items can serve as a source of *collateral information* for more accurately calibrating another targeted, dimensionally distinct and homogeneous subset of test items. Previous research has suggested a methodology by which variables external to the test can assist in the estimation of item parameters (Mislevy, 1987). The goal of this investigation was to determine whether items measuring distinct but correlated latent dimensions to that measured by the targeted set of items to be calibrated can likewise help produce better item parameter estimates.

Previous dimensionality analyses of Law School Admission Test (LSAT) forms suggest that the forms are well-approximated by a two dimensional structure, with items from the Analytical Reasoning (AR) section measuring a distinct ability from that measured collectively by the Reading Comprehension (RC) and Logical Reasoning (LR) sections (Ackerman, 1994; Camilli, Wang, & Fesq, 1992; Camilli, Wang & Fesq, 1995; De Champlain, 1995; De Champlain, 1996; Douglas, Kim, Roussos, Stout & Zhang, 1999; Stout, Habing, Douglas, Kim, Roussos, & Zhang, 1996; Wilson & Powers, 1994). Because of this finding, the Law School Admission Council (LSAC), as part of their investigation of a computerized version of the LSAT, could consider the possibility of reporting separate scores for the two dimensions. It may therefore be desirable to calibrate the RC items (along with the LR items) with respect to an ability dimension modeled as distinct from that of the AR items. In this context collateral information may be of particular benefit in item parameter estimation. Specifically, the use of collateral information from one section has the potential to allow the use of fewer test takers in estimating the item parameters of another targeted section. From the LSAC perspective, collateral information therefore finds an important potential application in calibrating pretest items. With respect to the anticipated LSAC Computerized Test (CT), this may translate to fewer administrations of pretest items to obtain equally precise item parameter estimates as compared to calibrating without collateral information.

In exploring the possible application of collateral information for item parameter estimation with the LSAT, this study had three specific objectives.

(1) to provide evidence that collateral information is useful in item parameter estimation;

(2) to evaluate the magnitude of collateral information improvements in pretest estimation in a pencil-and-paper (P&P) setting designed to be an analog of an LSAT CT setting; and

(3) to propose an easily-implemented methodology by which such collateral information can be utilized.

Several approaches were explored as possible ways of applying collateral information to improve pretest item parameter estimation. For purposes of investigation and demonstration, we assumed a parameter estimation scenario in which collateral information from an AR section is added to test taker responses from an RC section to calibrate pretest RC items to a desired level of accuracy. Although all results are obtained by way of data simulation, the parameters selected to generate the simulated data were based on estimates obtained using real data from test taker responses to operational items on the October 1992 administration of the LSAT, with parameters for the to-be-calibrated pretest items randomly selected from among those in the RC operational section. In this regard, the results of the simulation study should provide evidence of what is likely to occur when the methods are applied to real LSAT data. Parameters were estimated and data generated using the MD2PL multidimensional logistic model (Reckase, 1985):

$$P(\theta_1, \theta_2) = \frac{\exp(a_{i1}\theta_1 + a_{i2}\theta_2 + d_i)}{1 + \exp(a_{i1}\theta_1 + a_{i2} + d_i)}$$

where θ_1, θ_2 denote distinct abilities, a_{i1}, a_{i2} are corresponding discrimination parameters, and d_i a difficulty parameter. The program NOHARM (Fraser, 1988) can be used to obtain parameter estimates of the MD2PL function above. We may think of θ_1 and θ_2 as corresponding roughly to the abilities defined by the AR and RC content sections, although our methods do not make this assumption.

The effects of collateral information on item parameter estimation accuracy are evaluated by comparing the true item response functions (IRFs) of the pretest items with the IRFs obtained using the parameter estimates for the pretest items for conditions with and without collateral information. Note that because simulation data are used, the true IRFs are known to us. The mean absolute difference (our choice of metric for assessing IRF estimation accuracy) between curves is evaluated at 31 points on the θ scale, integrated across θ and weighting by the presumed standard normal density for the latent RC ability.

Note that the October 1992 LSAT administration consisted of 27 operational RC and 24 operational AR items. In a CT administration more precise estimates of ability could be obtained with respect to each of these ability dimensions because it would be possible to have more RC and AR sectional information than 27 RC and 24 AR P&P items can provide. In particular, LSAC could consider having each sectional CT score be as reliable as the current P&P overall LSAT test score. In close consultation with research scientists at LSAC, it was determined that one possible P&P analog (i.e., an analog in the sense of having the same amount of measurement precision) of such a CT test would consist of approximately 80 RC and 80 AR operational items. However, due to limitations of currently available item calibration software, it was sometimes necessary to restrict the number of operational items somewhat below this number. Fortunately, as shown below, this did not have a meaningful effect on the ultimate outcome, as the smaller number of operational items actually used appeared to reach an asymptotic condition.

Three approaches were explored for using collateral information from AR items to assist in calibrating RC pretest items. These methods are summarized in Methods 1, 2, and 3 below. A key challenge to successfully applying collateral information from AR operational items to improve RC pretest item estimation accuracy is to do so in a way that does not bias the RC item parameter estimates away from the RC direction and toward the AR measurement direction. Consequently, the amount of information provided by the AR items as they impact RC item parameter estimation in the RC ability direction should be reduced relative to the AR ability direction.

Method 1

The basic setting is a sample of test takers producing operational (RC, AR) item response data with a subset of the test takers also responding to a set of pretest RC items needing to be calibrated. The method splits into two stages. Stage 1 uses the operational (RC,AR) test taker response data to derive an AR collateral-information-enhanced θ_{RC} estimator, denoted by $\hat{\theta}_{RC}$. Stage 2 then uses $\hat{\theta}_{RC}$ as a known independent variable and the pretest RC item responses as the observed dependent variable in a standard logistic regression problem to estimate the pretest RC IRFs by a maximum likelihood estimation (MLE) procedure. Stage 2 is applied to the subset of the test takers who took both the operational items and the pretest RC items targeted for calibration. Stage 1 requires a more detailed description and is most easily described by breaking it down into several steps.

Step 1. Estimate the test structure of the operational (RC, AR) data using a two dimensional MD2PL model estimation procedure, our choice being NOHARM operated in a confirmatory simple structure mode. That is, all RC items are presumed to measure the same RC dimension and no other dimension, and similarly for the AR items.

Step 2. Using the item parameter estimates from Step 1 as the true item parameters and assuming a standard multinormal ability distribution, generate an artificial data set. Because LSAT dimensionality analyses have consistently supported a two-dimensional structure for the RC and AR items, the psychometric properties of this artificial data would be expected to approximately resemble the properties of the original LSAT data. The artificial data has the added advantage that the true θ_{RC} values are known for these data.

Step 3. Using the artificial data, obtain $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$ for the simulated test takers in these data by running BILOG (the marginal MLE based procedure of Mislevy & Bock, 1989) separately on the artificial AR and RC data sets.

Step 4. Using θ_{RC} as the dependent variable and using $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$ as the independent variables, solve for the multiple linear regression equation relating θ_{RC} to $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$.

Step 5. Assuming this regression equation closely approximates the relation of θ_{RC} to $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$ in the original LSAT data, this formula is applied to the operational (RC, AR) data. BILOG is run separately on the RC and AR operational data to obtain $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$ for all the test takers. Then the $\hat{\theta}_{RC}$ and $\hat{\theta}_{AR}$ values are substituted into the regression equation to obtain the collateral-information-enhanced $\hat{\theta}_{RC}$ values for all the test takers (actually, the only values needed are those for the test takers who took the pretest RC items). As mentioned above, these collateral-information-enhanced θ_{RC} ability estimates are then used in Stage 2 to estimate the IRFs for the pretest RC items.

Study 1

Study 1 was then conducted to evaluate the potential effectiveness of Method 1. To evaluate the method, it was applied to a simulated data set. To obtain realistic item parameters for the simulated data, NOHARM was applied in a two-dimensional exploratory mode (with uncorrelated abilities) to a real LSAT data set consisting of 5,000 test taker responses to 24 AR and 27 RC operational items from the October 1992 LSAT administration. The item parameter estimates from this NOHARM analysis were then used as the true item parameters for generating the simulated data to be used in the evaluation of the method. Using these item parameters for the simulated 24 AR and 27 RC operational items, using a random sample of 10 of the RC item parameters for the simulated RC pretest items, and assuming a standard bivariate normal ability distribution with uncorrelated abilities, a simulated data set was generated with 5,000 test takers taking the simulated operational items and 1,000 of these test takers also taking the 10 simulated pretest RC items. In this manner the true pretest RC IRFs for the data set were in fact known to us and thus the performance of Stage 2 of the method has the potential to be evaluated. Moreover, the true RC abilities for the 5,000 operational-item-taking test takers were also known to us. The RC ability direction was determined by taking the estimated measurement directions in the (θ_1,θ_2) plane for all the simulated RC items. This amounted to equating θ_{RC} to the appropriate weighted combination of θ_1 and θ_2. Hence the performance of the Stage 1 collateral-information-enhanced RC ability estimation for the 5,000 test takers taking the operational items has the potential to be evaluated. In particular the accuracy of using collateral information to estimate RC abilities can be compared with the ordinary BILOG estimated RC abilities using only test taker responses to the operational RC items.

An evaluation of the collateral-information-enhanced RC ability estimates in relation to the true θ_{RC} values was first conducted because for the Stage 2 MLE procedure to be effective, the collateral information enhanced θ_{RC} estimates must provide an improvement over the ordinary noncollateral information θ_{RC} estimates. Comparisons were made both with respect to the ability estimation standard errors and bias occurring locally at several locations in the θ_{RC}, θ_{AR} plane. Results demonstrated noticeably reduced estimation standard errors over the entire plane. Although serious θ_{RC} ability estimation bias occurred at certain locations, thus reducing or negating θ_{RC} ability estimation improved accuracy as assessed by mean square error. Consequently, until a method is found to control for this bias, it was deemed appropriate to consider other methods that bypass the θ_{RC} collateral information enhanced ability estimation stage of this method. However, it is planned in future research to use nonlinear regression of θ_{RC} on $(\hat{\theta}_{RC}, \hat{\theta}_{AR})$ in an attempt to reduce the RC ability estimation bias to an acceptable level, thereby achieving reduced mean square θ_{RC} ability estimation error locally in the $(\theta_{RC}, \theta_{AR})$ plane.

Method 2

This method bypasses the ability estimation stage entirely, although it still presumes two-dimensional IRF estimation capacity, such as provided by NOHARM or TESTFACT (Bock, Gibbons & Muraki, 1988). The first step is a two-dimensional calibration of the operational (RC,AR) items. In its usual exploratory mode, NOHARM provides a solution in which each item loads to varying degrees on a specified number of underlying common factors. With NOHARM there also exists the capacity to fit confirmatory factor models—for example, each subset of items of the multidimensional model that is presumed to measure a common dimension can be constrained so as to measure *only* that dimension (e.g., RC items can be presumed to measure only in the RC direction, and similarly for the AR items). A multidimensional test will be said to exhibit "simple structure" when its items can be partitioned into dimensionally homogeneous clusters. As referenced above, previous dimensionality studies of the LSAT suggest that such simple structure in a two-dimensional solution appears to be approximately satisfied. Consequently, a multidimensional NOHARM approach was attempted using both a confirmatory (assuming partial simple structure) as well as an exploratory (not assuming any simple structure) method.

In the exploratory method, MD2PL item parameter estimates for RC and AR items are computed for all of the items without any simple structure constraints imposed. An RC measurement direction is defined as the composite direction in the θ_1, θ_2 plane determined by averaging the estimated RC item measurement directions. (Note that the θ_1 found in NOHARM is not presumed to be θ_{RC} when simple structure is not imposed on the directions of item measurement.) Discrimination parameter estimates can be determined for all items (AR as well as RC) with respect to the RC measurement direction by geometrically projecting the discrimination parameter vector for the items onto that direction. We omit the mathematical details of this projection. However, its role is to model AR item functioning in the RC direction, thus allowing us to capture the collateral information provided by the AR items in the RC direction. Collateral information is obtained from the fact that both AR and RC items are calibrated simultaneously using a multidimensional estimation method and thus both supply information about each of the latent abilities underlying the solution. Segall

(1996), for example, demonstrates the usefulness of multidimensional estimation in increasing measurement precision in the context of multidimensional computerized adaptive testing.

In the confirmatory method, the RC items are assumed to measure only in the RC direction (a simple structure assumption) while the AR items' measurement directions are still allowed to vary and are thus determined by the usual exploratory factor analysis methods. The measurement directions for the AR items are allowed to vary because the goal of the analysis is to find out the degree of discrimination of each AR item in the RC direction. The factor pattern is displayed as in Figure 1. Note that in this case the RC items are constrained to measure a single common dimension, while the AR items are free to load both on the RC ability, as well as an orthogonal dimension. In this way, whatever information the AR items contain with respect to RC ability is automatically taken into account in calibrating the RC pretest items.

Item	θ_1	θ_2
1	1	0
2	1	0
3	1	0
.	1	0
.	1	0
27	1	0
28	1	1
29	1	1
.	1	1
.	1	1
50	1	1
51	1	1

FIGURE 1. *Factor pattern*

Study 2

The simulated data used to evaluate the exploratory method are generated in the same way as the data generated for Study 1. To evaluate the confirmatory method, simulated data are generated assuming an (AR,RC) simple structure with the simulated AR and RC items measuring dimensionally distinct but correlated abilities. The parameters used to generate data conforming to this structure were obtained by applying a NOHARM simple structure solution to the operational AR and RC sections of the October 1992 LSAT administration. The item parameters for the simulated pretest RC items were randomly selected from among the parameter estimates obtained for the RC operational items. In a manner similar to Study 1, the simulated abilities were generated from a standard bivariate normal distribution with a correlation of 0.7 between the two abilities.

The results for both variations of Method 2 suggested no discernible effect from collateral information in improving the simulated RC pretest item parameter estimates. In an attempt to determine whether such results were an artifact of the methodology used or an indication that useful collateral information did not exist, we tried several modifications of this basic setting to thoroughly assess NOHARM's capacity to make valid use of the available collateral information. For example, a strictly unidimensional NOHARM analysis was conducted by adding (noncollateral) information by merely increasing the simulated RC section length. In essence, this approach replaces the AR items by additional RC items known to measure the same ability dimension. Analyses were carried out both in the short sectional length condition (in which 24 RC items are added to the original 27 RC items) as well as in the (60,60) case, which is the maximum possible because of NOHARM operating constraints. The obtained results (in conjunction with those positive results to be described shortly under Method 3) suggested that our lack of positive results when using NOHARM were likely a result of an inadequate performance by NOHARM rather than a lack of collateral information. The reasons for NOHARM's inability to utilize collateral information from the simulated AR items were left as an issue for future study.

Method 3

One solution to the problems incurred in using NOHARM may be the development or use of alternative software. Limited by the fact that developing and coding effective two-dimensional estimation software is a challenging and time-consuming task, we turned to the use of alternative software. We devised a way to coerce BILOG into calibrating the AR operational items in the composite direction of the RC items. The first step required unidimensional BILOG estimation of the operational RC item parameters. The next step required repeatedly running BILOG on all 27 operational RC items together with one AR item at a time (24 separate runs of BILOG). In each of these BILOG runs, the RC item parameter estimates were fixed to be their values from the first BILOG run on the RC items alone (an option available with BILOG is the fixing of some of the IRF parameters). Consequently, each AR item is treated by BILOG as though it were an RC item; and, in particular, the estimated discrimination parameters for the AR items give the degree to which these items discriminate in the direction of best RC measurement. The final step involved fixing the estimates of all the operational items (using the estimates from the first BILOG run for the RC items and the estimates from the 24 BILOG runs for the 24 AR items) and using BILOG to estimate parameters for five RC pretest items.

Study 3

As was done in evaluating the confirmatory variation of Method 2, the simulated data used to evaluate this method are generated assuming an (AR,RC) simple structure with the simulated AR and RC items measuring dimensionally distinct but correlated abilities. The parameters used to generate the data were obtained (as in Method 2) by applying a NOHARM simple structure solution to the operational AR and RC sections of the October 1992 LSAT administration. Because LSAC is currently considering pretesting items in five-item sets, the number of simulated pretest RC items to be calibrated was set at five to add a further degree of realism to the study. The item parameters used to simulate the data for the targeted pretest RC items were randomly selected from among the parameter estimates obtained for the RC operational items. As in Study 2, the simulated abilities were generated from a standard bivariate normal distribution with a correlation of 0.7 between the two abilities.

Two types of data structure cases were simulated for the operational item responses. In the first case, 2,500 simulated test takers were generated, each of whom responded to all 51 simulated (RC,AR) operational items. In a second case, a scenario more realistic from the CT perspective was simulated in which operational items have varying exposure rates. In this case, one third of the operational items were responded to by 1,250 test takers, another third were responded to by 2,750 test takers, and the remaining third were responded to by 4,250 simulated test takers. In this latter case, it could be assumed that the accuracy of the operational item parameter estimates varied across items, as would occur with CT.

In both cases, additional data were generated for simulated test takers who took all 51 of the operational and all five of the pretest items. The simulated sample sizes for these additional item responses were 150, 200, 250, 500, 750, and 1,000. (Consultation with LSAC research scientists suggested that 1,000 test takers per pretest item may be a reasonable upper bound for the proposed LSAT CT).

Two conditions of collateral information were also studied. In the "no collateral information" condition, only the responses for the 27 simulated RC operational items (and their accompanying fixed value IRF estimates) along with the 5 simulated RC pretest items are used. In the "collateral" condition the 24 simulated AR items (and their fixed value one-at-a-time BILOG IRF estimates) are added to the calibration. The accuracy of parameter estimates for the 5 pretest RC items is evaluated by comparing the corresponding estimated IRFs with the true IRFs (known to us because we are working with simulated data) used to generate the items, in the manner described earlier. Results are displayed in Figures 2 and 3 for the case where 2,500 test takers were simulated for all operational items and for the case where 1,250, 2,750, or 4,250 test takers were each simulated for one third of the operational items, respectively. A total of 6 replications were conducted for both the "no collateral" and "collateral" conditions at each sample size level. In both figures the mean absolute error between true and estimated IRFs (across the five pretest items and averaged over the six replications) are presented. Negative exponential curves are fit separately for the "no collateral" (X) and "collateral" (O) conditions to the overall mean across replications at each sample size level in an effort to extract signal from noise. The distance between curves provides an estimate of the amount of error reduction that can be attributed to collateral information, as the lower curve in both figures corresponds to the collateral condition. For example, it appears from Figure 2 that the accuracy attainable with approximately 700 test takers when not using collateral information can be attained with about 600 test takers when using collateral information. This is easily seen by drawing a vertical line from 700 to the upper curve, then a horizontal line to the lower curve, then a vertical line back down to the horizontal axis and reading the estimated savings in number of test takers required, which is about 100. The difference between curves appears to be approximately the same across the conditions in which the number of simulated test takers per operational item was varied.

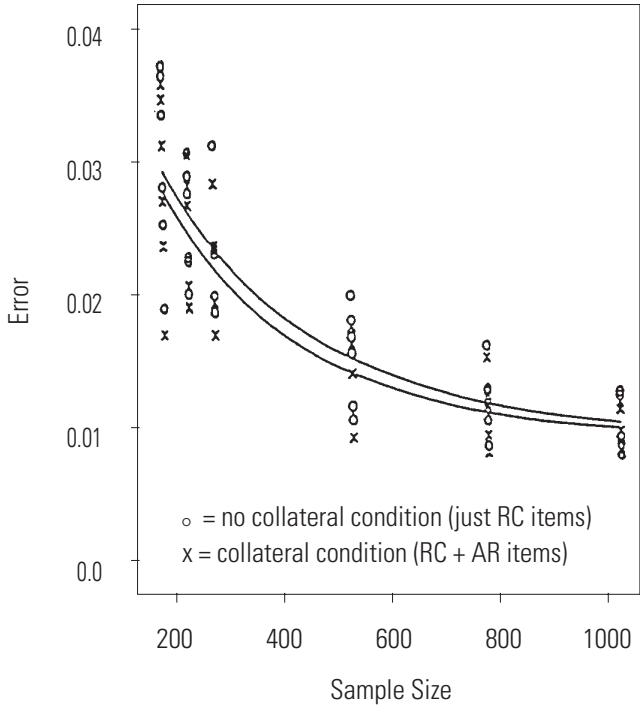

FIGURE 2. *Case of 2,500 test takers per operational item*

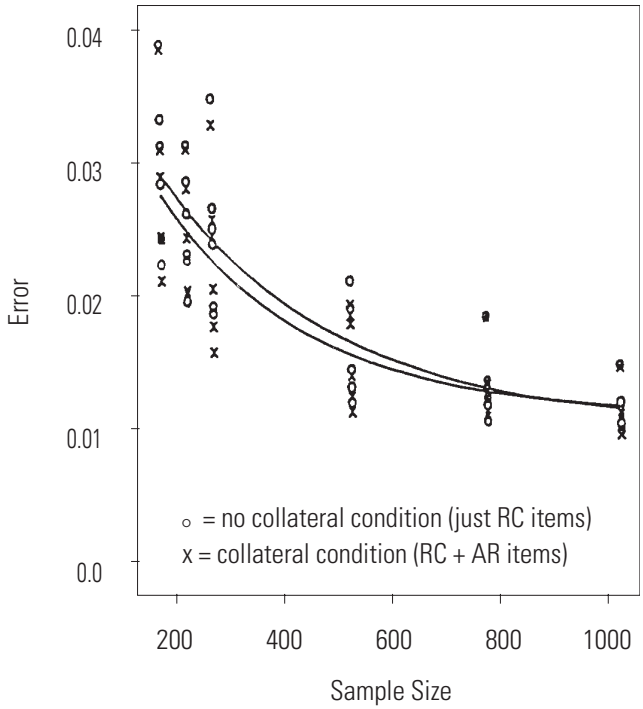

FIGURE 3. *Case of varying number of test takers per operational item*

The effects of collateral information were much less pronounced when the number of operational items was increased to emulate the level of measurement precision expected under an LSAT CT administration. That is, the approach just described was reconducted using 60 simulated RC and 60 simulated AR items (BILOG restrictions would not enable more operational items). Figures 4 and 5 display the means across the six replications at each sample size under the "no collateral" and "collateral" conditions. In both the fixed and variable operational sample size cases, there appears to be virtually no effect as a result of collateral information. It appears that 60 well-estimated RC items provide all the information we need in estimating pretest RC IRFs, keeping in mind that there are other influential sources of estimation error such as the inherent randomness of test taker responding. Because these results with 60 items each of AR and RC showed little if any gain from collateral information, it is clear the same result would occur if 80 items were simulated for each of AR and RC. Thus the 60-item limitation turned out not to be a problem for generalizing our results to the case of using the desired larger number of items. Because of the effectiveness of BILOG in capturing useful collateral information, we made a preliminary effort to write a marginal MLE BILOG-like program following a prototype provided by Baker (1992). However, we abandoned this effort when it was found that its errors of IRF estimation were 50% higher than with the BILOG program.

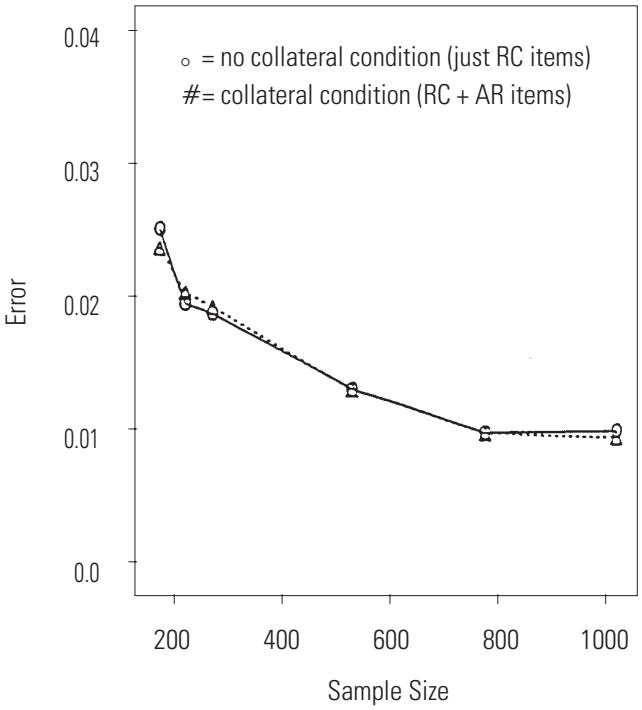

FIGURE 4. *Case of 2,500 test takers per operational item*

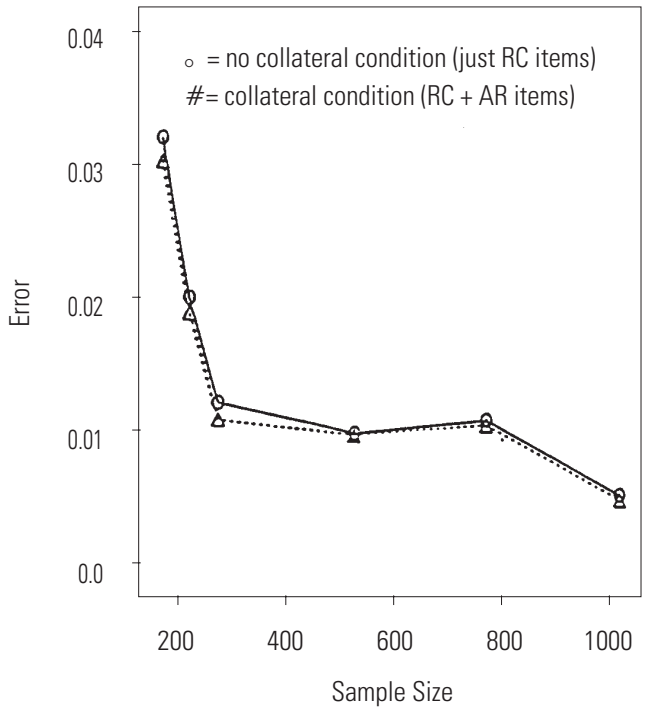

FIGURE 5. *Case of varying number of test takers per operational item*

Summary and Conclusion

The results of Method 3 appear to be quite convincing in their demonstration of the potential increase in accuracy of item parameter estimation through collateral information. Unfortunately, for purposes of LSAT CT administration as currently planned, the substantial improvement in ability estimation seems to offset any advantage that could be gained through collateral information, at least when assuming the two-dimensional structure of the LSAT considered here. Perhaps the best demonstration of collateral information effects will occur in settings where smaller numbers of operational items are used to define a dimension according to which new items must be calibrated. For example, we conjecture that in cases like (10 RC, 10 AR) or (15 RC, 15 AR) operational items that the improvements in test taker pretest sample sizes needed should be quite dramatic, even when compared with the Figures 2 and 3 (27 RC, 24 AR) operational items case. Although the focus of this study precluded a more detailed dimensionality analysis of the LSAT, if item parameter estimation of LSAT forms with respect to a larger number of dimensions is ever deemed useful, collateral information may be of greater value. In the case detailed in this paper, the use of only two dimensions resulted in a large number of operational items per dimension, large enough that it appears to put us in the asymptotic situation where using an even larger number of operational items in the calibrating of pretest items provides via collateral information little if any added benefit.

For reasons that were not explored in this study, the estimation approach employed by the BILOG program seems to make more effective use of collateral information than that employed by the multidimensional estimation program NOHARM. Development of software emulating the algorithm of BILOG but permitting the fit of multidimensional structures such as in Figure 1 would appear to be a useful goal of future work. Because the operational items are fixed at their previously estimated values (i.e., no multidimensional estimation is required), development of such software would appear to be reasonably straightforward.

References

Ackerman, T. (1994, April). *Graphical representation of multidimensional IRT analysis*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques.* New York: Marcel Dekker.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement, 12(3)*, 261-280.

Camilli, G., Wang, M. M., & Fesq, J. (1992*). The effects of dimensionality on true score conversion tables for the Law School Admission Test* (Research Report 92-01). Newtown, PA: Law School Admission Council.

Camilli, G., Wang, M. M., & Fesq, J. (1995). The effects of dimensionality on equating the Law School Admission Test. *Journal of Educational Measurement, 32*, 79-96.

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.

DeChamplain, A. F. (1996). The effect of multidimensionality on IRT true-score equating for subgroups of examinees. *Journal of Educational Measurement, 33,* 181-201.

Douglas, J., Kim, H. R., Roussos, L., Stout, W., Zhang, J. (1999). LSAT Dimensionality Analysis for the December 1991, June 1992, and October 1992 Administrations (Statistical Report 95-05) Newtown, PA: Law School Admission Council.

Fraser, C. (1988). *NOHARM: A computer program for fitting both unidimensional and multidimensional normal ogive models of latent trait theory*. Center for Behavioral Studies, The University of New England Armidale, New South Wales, Australia.

Mislevy, R. J. (1987). Exploiting auxiliary information about examinees in the estimation of item parameters. *Applied Psychological Measurement, 11(1)*, 81-91.

Mislevy, R. J., & Block, R. D. (1989). *BILOG: Item analysis and test scoring with binary logistic models* [Computer program]. Mooresville, IN: Scientific Software.

Reckase, M. (1985). The difficulty of items that measure more than one ability. *Applied Psychological Measurement, 9(4)*, 401-412.

Segall, D. (1996). Multidimensional adaptive testing. *Psychometrika, 61(2)*, 331-354.

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

Wilson, K. M., & Powers, D. E. (1994). *Factors in performance on the Law School Admission Test* (Statistical Report 93-04). Newtown, PA: Law School Admission Council.