ONLINE ITEM PARAMETER RECALIBRATION: APPLICATION OF MISSING DATA TREATMENTS TO OVERCOME THE EFFECTS OF SPARSE DATA CONDITIONS IN A COMPUTERIZED ADAPTIVE VERSION OF THE MCAT

J. Christine Harmes

Jeffrey D. Kromrey

Cynthia G. Parshall

University of South Florida

Report submitted to the Association of American Medical Colleges, Section for the MCAT,
October 2001

Background and Rationale

Migration of standardized tests from traditional paper-and-pencil administration to computerized adaptive administration offers many potential benefits, including increased measurement efficiency, immediate scoring, and more frequent administration dates. Along with these benefits come potential difficulties. While many testing programs conduct initial computerized adaptive tests (CATs) using item parameter estimates from previous paper-and-pencil administrations, a more sound practice would suggest that calibration from computer administration is best (Haynie & Way, 1995; Ito & Sykes, 1994). A testing program could elect to begin operational CAT administration using paper-and-pencil calibrations and then recalibrate when sufficient online data have been collected. Recalibration is also recommended periodically in order to address possible scale drift (Stocking, 1988). Problems that may affect parameter accuracy if items are not recalibrated include a mode effect (i.e., differences between paper-based and computer-based administration such as item ordering, item review, and context), and potential cognitive differences between the two modes (Parshall, 1998).

The major sources of difficulty in recalibrating parameters for a test that is being administered as a CAT lie in the restriction of ability range and sparseness of the data matrices, i.e., missing data. Each of these difficulties is discussed further.

An optimal distribution of examinee ability for calibrating item parameters is a broad, possibly uniform, distribution (Stocking, 1990). Since the CAT is designed to maximize efficiency, examinees are generally given items that are targeted to provide the most information at their estimated ability level (depending upon the item selection algorithm). This results in the more difficult items only being given to higher-ability examinees, and the less difficult items only being administered to lower-ability examinees. This produces a restriction of the ability range available for item parameter calibration, and has the potential to affect calibration accuracy (Haynie & Way, 1995; Ito & Sykes, 1994; Parshall, 1998).

CAT administration also results in a sparse data matrix to be presented to a calibration program (e.g., BILOG). This sparseness is the result of the size of the pool in relation to average test length, and to the use of targeted item selection. For test security purposes, an item pool typically contains far more items than are administered to any single examinee. Recommendations for the size of the CAT item pool are offered in terms of the total number of items available compared to the length of the fixed CAT exam (a typical test's length for an adaptively administered exam might be about half the length of the paper and pencil test form it replaces). Way (1998) suggests an item pool size for certification testing consisting of 6-8 times as many items as would be on an average CAT administered from that pool, while Stocking (1994) suggests a pool size of up to 12 times the average adaptive test length.

When the data are recalibrated after an adaptive administration, the examinee response records contain many more items that were not presented than items that were administered to each test taker. For example, in a fixed-length, 30-item test with a pool containing 360 items, each examinee record would include scored responses to 30 items and missing data on 330 items. Further, the examinees who take a CAT tend to have relatively few items in common, which is very different from fixed-form paper-and-pencil testing. This lack of item overlap will increase the problem of sparseness.

Sparseness in the calibration data set has the potential to affect the quality of the item parameter estimates (Haynie & Way, 1995; Hsu, Thompson, & Chen, 1998). If the results of a sparse data calibration are compared to the calibration results for a full data matrix from paper-

and-pencil administration, approximately 10 times the number of examinee response records may be needed if the standard approach to calibration is used (Hsu, Thompson, & Chen, 1998).

The problem of sparseness in the data matrix is essentially a problem of missing data. In the case of adaptive test administration, the cause of the missing data is systematic (i.e., non-ignorable nonresponses). In a CAT, items are selected for administration based on an algorithm that usually relies on the estimate of the examinee's ability along with the maximum information each item provides at every ability level. Thus, the reason for the sparseness in the data matrix is nonrandom. This kind of nonrandom missingness is related to the restricted range problem. Research on strategies for dealing with nonrandomly missing data (Kromrey & Hines, 1994; Little & Rubin, 1987) offers possible solutions for application to the CAT sparseness problem. A strategy that might be useful is multiple imputation. The multiple imputation technique involves generation of imputed values for missing data from their posterior distribution based on existing, observed data (Thomas & Gan, 1997). More than one value is imputed for each missing element (producing multiple data matrices with different imputed values for the missing data), allowing for a more accurate estimation of the variance of the estimates (Little & Rubin, 1987).

Multiple Imputation

Multiple imputation is a technique for "filling in" missing values in a dataset. Using the information provided by the observed data, a series of calculations can be made to impute what the missing values might have been, had they been observed. Multiple imputation is an extension of single imputation.

Single imputation is a common technique employed by researchers and available in data analysis software packages (e.g., SAS). In general, single imputation offers the advantage of allowing complete data analysis methods to be used, and it requires much less work to impute each missing value only once (Rubin, 1987). A common example of single imputation is replacing missing values with the mean observed value for that variable. This technique is often referred to as mean imputation. A problem with mean imputation is that it distorts the distribution of that variable once the means have been used to replace missing values, and variance is underestimated (Little & Rubin, 1987). Another example of single imputation is hot deck imputation, in which individual missing values are replaced by values from similar responding units drawn from an estimated distribution (Little & Rubin, 1987). This technique does not distort the distribution in the way that mean imputation does, but it does not represent uncertainty as does multiple imputation.

With multiple imputation, two or more acceptable values are calculated for each missing data point, representing the distribution of possible values (Rubin, 1987). This strategy was originally proposed by Rubin (1977; 1978) as a technique for dealing with missing data in surveys. For each missing value, a vector of imputed values is created. The number of imputations, or *m*, is generally from two to ten (Rubin, 1987; Schaefer, 2001). If *m* were equal to ten, for each missing value, ten acceptable values would be imputed, thus allowing for the creation of ten possible full data sets. Each complete data set can then be analyzed with standard procedures. Advantages of multiple imputation over single imputation include the fact that the practice of randomly drawing imputed values from the distribution of possible values increases estimation efficiency. Additionally, the fact that imputed values are repeatedly being drawn at random allows for the straightforward combination of inferences drawn from analysis of the *m* complete data sets. Disadvantages of multiple imputation are the fact that it requires more effort than single imputation, it requires more computer storage, and the effort required to analyze the complete datasets is multiplied.

Missing Data and CAT/IRT

Difficulties encountered in attempting to recalibrate item parameter estimates for a project on the CAT-ASVAB (Parshall, Kromrey, Harmes, & Sentovich, 2001) for the Department of Manpower Defense Contractors (DMDC), have confirmed the problems of calibrating with a sparse data matrix. When item response files were presented to BILOG, the program was not able to successfully calibrate all items. A subset of items had to be drawn that had the highest overlap rates. BILOG was then able to calibrate this subset of commonly occurring items.

Much of the work that has been done in the CAT field dealing with problems with item calibration due to sparse data have focused on pretest item calibration (e.g., Ban, et. al, 2001; Haynie & Way, 1995; Stocking, 1988; Wainer & Mislevy, 1990). In the case of pretest item calibration, the goal is to administer items that are not yet operational, but are being tested for inclusion in a subsequent form of the exam. In order to determine whether or not the items should be used in the future, and to prepare them for operational use, they must be calibrated. Several techniques have been suggested for calibrating pretest items.

The PIC program (Spray, 1995) allows for the calibration of a set of pretest items, given the item parameters of a set of operational test items that were administered at the same time. The program treats the item parameters of the operational items as fixed, and only calibrates the pretest items. This same logic is applied in many of the research studies described below.

Stocking (1988) investigated two methods (Method A and Method B) for online pretest item calibration. In Method A, operational items were used to compute ability estimates for simulated examinees. Computerized fixed testing (i.e., nonadaptive) was simulated for 50 pretest items. Method B involved the use of anchor items. A set of 25 anchor items was selected from a 100 item pool. Fixed testing was simulated for five anchor items and five pretest items with each examinee. This resulted in 1500 responses for each pretest item and 3000 responses for each anchor item. Five rounds of simulation were conducted. Results at the end of all rounds showed Method B to have smaller bias than Method A. At the low and high ends of the ability continuum, root mean-squared error (RMSE) was similar across the two methods. In the middle ability range, Method B showed a smaller RMSE.

Wainer & Mislevy (1990) suggested a method of “seeding” pretest items into CAT exams. Each examinee could be given one or two new items near the beginning of his or her test. These pretest items would not contribute toward proficiency estimation. This would allow for the calibration of these new items along with the operational items with which they were administered. Calibration of the new items could be carried out by MML estimation with an EM algorithm with a single E step and a single M step. The E step would use the parameter estimates from the operational items, and the M step would only maximize the likelihood functions for the new items.

Ban, et. al (2001) investigated the use of several techniques for pretest item calibration. Among these was ML estimation with the EM algorithm. In this case, the algorithm was written to compensate for missing item response data as the unknown parameter (as opposed to ability as the unknown parameter as in BILOG). To obtain true item parameters for the study, real response data was used to calibrate (3PL) 940 items. For the study, 240 of these items were assigned to be pretest items, 100 were used as anchor items, and 600 were used as the operation item pool. An initial set of 3000 examinee responses were generated for the 600 operational items. These were used to calibrate item parameter estimates for these items. A 30-item fixed-length adaptive test was then simulated for 12,000 examinees from a normal distribution of ability. Each simulee was also given ten pretest items and ten anchor items. The simulations were repeated 100 times. The resulting item response data matrices were sparse in for operational, pretest, and anchor items. Pretest item parameters from each method (the EM algorithm with one iteration (OEM), the EM algorithm with multiple iterations (MEM), and Stocking's Method B) were compared in terms of bias, standard error, and root mean squared error (RMSE) for each item parameter. The MEM method resulted in the smallest overall error across the 240 items.

Recalibrating operational items based on CAT data has been suggested as an important step (e.g., Kingsbury & Houser, 1993) but fewer studies have focused specifically on this problem. Haynie & Way (1995) used real and simulated data to compare item calibrations for both pretest items and operational CAT items. The original difficulty estimates were based on a sample of 1500 examinees while the real data came from 25,931 examinees. The simulated data was generated for 25,000 examinees using a 1,480 item pool for the CAT portion and a 200 item pool for the pretest portion. The 1PL model was used to calibrate item difficulties. They found greater differences between the original and recalibrated difficulty parameters in the real data than in the simulated data for both the CAT and pretest items. They attributed this difference to context effects.

Purpose

In order to ensure a smooth transition of standardized testing programs to CAT administration and to allow for recalibration of existing CAT programs, issues related to the quality of the item parameter estimates should be addressed. This study draws possible solutions from the literature on missing data and applies those approaches to IRT and CAT calibration problems. Specifically, the technique of multiple imputation for treating missing data was applied to the problem of item recalibration in a computerized adaptive testing setting.

Implications

In the process of investigating issues related to conversion of the existing paper-and-pencil testing programs to CAT administration, examining the seriousness of the problem of sparseness and determining what kind of errors in calibration it might cause is an important issue. Item parameter accuracy is critical in a CAT, as every aspect of the testing program is based on these parameters, from information functions, to interactive selection of items for examinees, to computation of final ability estimates. If these parameters are inaccurate or unstable, the integrity of the computerized testing program is in jeopardy, and the validity of the inferences made from these test scores is threatened.

Method

This study addressed the seriousness of the effect of sparseness on the recovery of item parameters and the efficiency of alternative procedures, through a simulation study.

Data Source

Data used in the study were from the Medical College Admissions Test (MCAT) Biological Sciences section. The MCAT is a standardized test designed to predict success in medical school. It is comprised of four sections: (a) Biological Sciences, (b) Physical Sciences,

(c) Verbal Reasoning, and (d) Writing. The objective sections of the test are reported to have a reliability estimate of approximately 0.85. The Biological Sciences section consists of multiple-choice items, both passage-based and discrete. In its current paper-and-pencil based format, this section consists of 77 multiple-choice items. The Biological Sciences section covers topics in biology and organic chemistry (Association of American Medical Colleges, 1998). For the calibration of item parameters that were referred to as the "true" parameter estimates, data from three spring administrations of the MCAT were used. The data set consisted of results from six forms of the Biological Sciences test, with 3000 to 6000 examinee responses for each form. Only the passage-based items were used in the simulation, yielding a dataset of 312 items.

Generation of "True" Item Parameters

Examinee response data from paper-and-pencil administrations were used to calibrate the items from each of the six test forms. This calibration resulted in item parameter estimates that were all on the same IRT scale. These items were then used to construct a 312 item pool. The parameter estimates for the items in the pool were regarded as baseline data, and became the "true" item parameters for purposes of comparison with those parameters calibrated based on the simulated data.

CAT Simulation

An adaptive test administration was then simulated, which resulted in a sparse matrix of simulated examinee response data.

Computerized test administration was simulated under four conditions: whole pool, random, no exposure control, and Sympson-Hetter exposure control. Whole pool administration served as the control condition in which each simulated examinee was given every item in the pool. This condition resulted in a full data matrix obtained under simulated administration. The random administration condition resulted in randomly missing data, and served as a baseline comparison. The two types of CAT administration (no exposure control and Sympson-Hetter exposure control) were simulated as the conditions under investigation.

Generation of Examinees

Examinee abilities were generated to represent a normal distribution. This provides a simulation condition that is similar to actual examinee populations, and allowed for an accurate representation of the level of sparseness in the data matrix that can be expected with a CAT administration.

Number of Examinees

The minimum number of examinees needed for 3-PL item parameter calibration is approximately 1, 000 (Wainer & Mislevy, 1990). Hsu, Thompson, and Chen (1998) recommend that approximately 10 times the number of examinees will be needed to calibrate with online data in order to have a level of accuracy that is comparable to calibration with a full dataset from paper-and-pencil administration. Based upon this recommendation, 10,000 examinees were simulated in the study.

Test Length

A major factor in the degree of sparseness in the data matrix should be the ratio of test length to item pool size. Based on recommendations of item pool size (Stocking, 1994; Way, 1998) ranging from 6 to 12 times the average adaptive test length, a test length of 36 items was chosen. This falls in the middle of the recommended range, representing a test length of approximately 1/9 the size of the item pool. In order to maintain the ratio of biology and organic chemistry items found in paper-and-pencil testing, five passages from biology were administered with four from chemistry; four items were administered per passage.

Exposure Control

In a CAT, there is a tendency for the items that provide the most information to be given to a large number of examines. This can lead to problems of item over-exposure (i.e., items that are seen by too many examinees). Several methods have been developed to deal with this problem. One of the most widely used in operational testing is the Sympson-Hetter (S-H) method (Sympson & Hetter, 1985). This method was used to control item exposure in one of the CAT conditions of the simulated test administration.

Missing Data Treatment with Multiple Imputation

Multiple imputation was used to fill in the missing data and generate complete datasets for recalibration. A set of 10 imputations was generated within each replication. Each imputed dataset was created by using the simulated examinee's estimated ability along with the parameters of each item to probabilitistically impute the examinee's score on all items that were not presented in the simulated exam.

Once the 10 imputations were made, they were used to create 10 complete datasets, each of which could then be analyzed with standard methods. To then combine the results, the following equations (from Schafer, 2001) were used.

$$\overline{Q}_j = \frac{1}{m} \sum \hat{Q}_{ij} \qquad (1)$$

where,

$\overline{Q}_j$ is the average of the values of a parameter estimate (such as *a, b*, or *c*) for the j^{th} item calculated from *m* complete data sets,

m = the number of imputation cycles,

$\hat{Q}_{ij}$ = the parameter estimate from the i^{th} imputation for the j^{th} item based on the data from one of the *m* complete datasets.

To compute the standard error of the average parameter estimate, the within imputation and between-imputation variances were first calculated and combined to find total variance. In Equation 2, the within-imputation variance is computed by using each estimate's error, U_{ij}, for each Q_{ij} The between-imputation variance was calculated by using Equation 3

$$\bar{U}_j = \frac{1}{m} \sum U_{ij} \qquad (2)$$

$$B_j = \frac{1}{m - 1} \sum (Q_{ij} - \bar{Q}_j)^2 \qquad (3)$$

Total variance was calculated with the following formula:

$$T_j = \bar{U}_j + \left(1 + \frac{1}{m}\right) B_j \qquad (4)$$

The square root of the value computed in Equation 4 provides the standard error of the averaged parameter value.

Evaluation of Item Parameter Estimates Across Replications

The entire set of test administration simulations and missing data treatment procedures was carried out 10 times across all conditions. Results from each trial were averaged for data analysis purposes. This repetition was done in order to reduce the impact of sampling error in the simulation (Robey & Barcikowski, 1992).

The item parameter estimates (*a, b, c*) resulting from each experimental condition were compared to the control conditions. The accuracy and stability of the item parameter estimates from the experimental conditions were evaluated based on their bias, standard error, and root mean-squared error (RMSE).

Statistical bias represents the difference between the average parameter estimate and the true value of the parameter. This was estimated as:

$$Bias(\lambda_{jk}) = 1/N \, \Sigma \, (\lambda_{jkn} - \lambda_{jk}), \qquad (5)$$

where,

λ_{jkn} = the j^th^ parameter estimate for the k^th^ item obtained from the *n*^th^ sample,

λ_{jk} = the population parameter, and the summation is over the N samples included in the simulation study.

The standard error of the estimate represents the difference between a parameter estimate and the value of the average parameter estimate. This was given by

$$SE(\lambda_{jk}) = \sqrt{\frac{\Sigma(\lambda_{jk} - \overline{\lambda}_{k})^2}{N}} \qquad (6)$$

where $\overline{\lambda}_{jk}$ = the average estimate of the j^th^ parameter for the k^th^ item, and the other elements are as defined above.

Estimation of RMSE represents total error, as it includes both bias and sampling error. It represents the average difference between a parameter estimate and the true value of the parameter. This was estimated as

$$RMSE((\lambda_{jk}) = \left[1/N \, \Sigma \, (\lambda_{jkn} - \lambda_{jk})^2\right]^{1/2}, \qquad (7)$$

where the elements are as defined above.

Limitations

The use of simulation as a method restricts the degree to which inferences can be made to operational testing. However, simulation makes a study such as this possible, which would otherwise be difficult, due to constrains of time, money, and availability of large numbers of participants. The use of simulation to generate examinee data provides the unique opportunity of conducting a study in which “truth” is known (e.g., true item parameters).

Results and Conclusions

Item parameter estimates calculated under each of the study conditions were compared to the “true” item parameter estimates calibrated from the original paper-and-pencil data. Results were examined in terms of the bias, standard errors of the estimates, and root mean-squared errors of the item parameters from the datasets treated with multiple imputation.

Variability Across Imputations

Within each replication, variance across the item parameter estimates resulting from each set of imputations was analyzed. The values for total variance in item parameter estimates averaged across the sets of imputations and across replications are given in Table 1.

Condition	TOTAL VARIANCE					
	a		*b*		*c*	
	AVG	MAX	AVG	MAX	AVG	MAX
No Exposure Control	0.00352 *(0.00253)*	0.03032	0.02164 *(0.03355)*	0.23968	0.00219 *(0.00201)*	0.00956
Random	0.00354 *(0.00273)*	0.27711	0.01879 *(0.02944)*	0.19857	0.00199 *(0.00181)*	0.00959
Sympson-Hetter	0.00352 *(0.00259)*	0.02260	0.02067 *(0.03261)*	0.26479	0.00212 *(0.00194)*	0.00975

Table 1. Total variance of parameter estimates across sets of imputations.

There is virtually no difference in the average total variance of the *a* parameter across the three CAT administration conditions. The random condition evidenced the highest maximum total variance (0.27711). Conversely, for the *b* parameter the random condition had the lowest average variance (0.01879) and substantially lower maximum variance (0.19857). Similarly, the random condition had the lowest average variance for the *c* parameter (0.00199). The Sympson-Hetter condition yielded the highest maximum variance for the *c* parameter (0.00975).

Bias

Statistical bias, the difference between the average parameter estimate and the true value of the parameter, was calculated for the *a, b*, and *c* parameters across the replications. Results are given in Table 2 and Figures 1-3. The calibration of results based upon the whole pool is included as a reference distribution.

Condition	BIAS					
	a		*b*		*c*	
	AVG	MAX	AVG	MAX	AVG	MAX
Whole Pool	0.03694 *(0.01873)*	0.12925	0.08048 *(0.10786)*	0.63513	0.0206 *(0.02378)*	0.11381
No Exposure Control	0.04400 *(0.01384)*	0.10675	0.06512 *(0.25523)*	0.95998	-0.02428 *(0.01789)*	0.02623
Random	0.06359 *(0.01862)*	0.14639	0.05635 *(0.26771)*	1.02035	-0.02770 *(0.01946)*	0.01192
Sympson-Hetter	0.04919 *(0.01453)*	0.11221	0.06176 *(0.25831)*	0.98674	-0.02500 *(0.01796)*	0.01989

Table 2. Bias of item parameter estimates across replications.

The random condition yielded the highest average amount of statistical bias for the *a* parameter (0.06359), while the whole pool condition had the lowest average bias (0.03694). The highest maximum bias was found in the random condition (0.14639). For the *b* parameter, the whole pool condition resulted in the highest average bias (0.08048), while the random condition showed the highest maximum bias (0.95998). All three of the CAT administration conditions evidenced greater variability in the *b* parameter biases (Figure 2). The three experimental conditions all resulted in slightly negative bias on the *c* parameter, with the random condition having the strongest negative bias (-0.02770). The highest value of bias for the *c* parameter was found in the whole pool condition.

Figure 4 shows a plot of the bias in *b* parameter estimates by true item difficulty. The three experimental conditions show nearly identical performance, with pronounced positive bias on the easier items and negative bias on the more difficult items. In contrast, the whole pool condition showed no relationship between the size or direction of the bias and the actual value of the parameter.

Standard error of the estimates

Standard errors of the estimates, which represents the difference between a parameter estimate and the value of the average parameter estimate, were calculated for the *a, b*, and *c* parameters across the replications. Results are given in Table 3 and Figures 5-7.

Condition	STANDARD ERROR					
	a		*b*		*c*	
	AVG	MAX	AVG	MAX	AVG	MAX
Whole Pool	0.03725 *(0.01493)*	0.14092	0.09668 *(0.06383)*	0.43742	0.02968 *(0.01426)*	0.07756
No Exposure Control	0.07554 *(0.02143)*	0.15206	0.08136 *(0.04935)*	0.29257	0.02435 *(0.00692)*	0.04664
Random	0.07750 *(0.02207)*	0.15556	0.08030 *(0.0499)*	0.28311	0.02376 *(0.00688)*	0.04700
Sympson-Hetter	0.07619 *(0.02147)*	0.15239	0.08127 *(0.04917)*	0.29167	0.02433 *(0.00699)*	0.04658

Table 3. Standard error of parameter estimates across replications.

As illustrated in Figure 5, the whole pool condition evidenced the lowest average standard error for the *a* parameter estimates, with the other conditions having nearly three times as much error. The standard errors for the *b* parameter estimates show very little difference across all four conditions, with six items showing exceptionally large standard errors under the whole pool condition. The typical standard error for the *c* parameter was smaller under all three of the CAT conditions, and further, showed less variability across standard errors than the whole pool condition (Figure 7).

Root mean-squared error

Root mean-squared error, the average difference between a parameter estimate and the true value of the parameter, was calculated for the *a, b*, and *c* parameters across the replications. Results are given in Table 3 and Figures 8-10.

Condition	RMSE					
	a		*b*		*c*	
	AVG	MAX	AVG	MAX	AVG	MAX
Whole Pool	0.00332 *(0.00268)*	0.02743	0.03148 *(0.05965)*	0.54449	0.00207 *(0.00231)*	0.01567
No Exposure Control	0.00829 *(0.00419)*	0.03059	0.07822 *(0.13189)*	1.02873	0.00155 *(0.00144)*	0.01151
Random	0.01088 *(0.00587)*	0.04563	0.08345 *(0.13829)*	1.12128	0.00176 *(0.00165)*	0.01117
Sympson-Hetter	0.00889 *(0.00452)*	0.03582	0.07934 *(0.13425)*	1.05873	0.00159 *(0.00144)*	0.01156

Table 3. RMSE of parameter estimates across replications.

All three CAT administration conditions yielded higher root mean-squared error for the *a* parameter than the whole pool administration, with the random condition having the highest average RMSE (0.01088). Similarly, all three CAT conditions revealed higher RMSE for the *b* parameter estimate, with many items showing exceptionally large values (greater than one). In contrast, the whole pool condition resulted in the highest RMSE for the *c* parameter, and also produced higher extreme values for specific items.

The results of the research must be interpreted within the context of its limitations. Only a single item pool was used, with a single test length and a single sample size. Additionally, only 10 replications were conducted. Within the framework of these limitations, however, certain patterns in the data merit consideration. The effect of sparseness depends upon the parameter being estimated. Overall bias was evident in the *a* and *c* parameters, but not in the item difficulty (*b*). Although the overall bias in *b* was negligible, the bias was related to the actual item difficulty. That is, the bias is such that easier items appear to be more difficult, and harder items appear to be easier. Surprisingly, differences across CAT item selection methods were negligible. The effect of sparseness on the parameter estimates appears to be unrelated to the degree to which missingness is correlated with examinee ability.

Overall the data suggest that the multiple imputation technique is a promising tool for the treatment of missing data in the context of item calibration in adaptive testing. The conditions examined in this study represented more than 88% missing data in the item response matrices. The relatively small levels of statistical bias and increases in sampling error underscore the potential of multiple imputation in applied testing programs.

References

Association of American Medical Colleges. (1998). *MCAT interpretive manual: A guide for understanding and using MCAT scores in admissions decisions*. Washington, DC: Author.

Ban, J., Hanson, B.A., Yi, Q., & Harris, D. (2001, April). *Data sparseness and online pretest calibration/scaling methods in CAT.* Paper presented at the annual meeting of the American Educational Research Association, Seattle.

Haynie, K.A. & Way, W.D. (1995, April). *An investigation of item calibration procedures for a computerized licensure examination.* Paper presented at a symposium entitled Computer Adaptive Testing, at the annual meeting of the National Council on Measurement in Education, San Francisco.

Hsu, Y., Thompson, T.D., & Chen, W. (1998, April). *CAT item calibration*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego.

Ito, K. & Sykes, R.C. (1994). *The effect of restricting ability distributions in the estimation of item difficulties: Implications for a CAT implementation.* Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Kingsbury, G.G., & Houser, R.L. (1993, Spring). Assessing the utility of item response models: Computerized adaptive testing. *Educational Measurement: Issues and Practice*, 21-27.

Kromrey, J.D. & Hines, C.V. (1994). Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments. *Educational and Psychological Measurement 54*(3), 573-593.

Little, R.J.A. & Rubin, D.B. (1987). *Statistical analysis with missing data*. New York: Wiley & Sons.

Parshall, C.G. (1998, September). *Item development and pretesting in a computer-based testing environment.* Paper presented at the ETS Sponsored Colloquium on *Computer-Based Testing: Building the Foundation for Future Assessments,* Philadelphia.

Parshall, C.G., Kromrey, J.D., Harmes, J.C., & Sentovich, C. (2001). *Nearest neighbors, simple strata, and probabilistic parameters: An empirical comparison of methods for item exposure control in CATs.* Paper presented at the annual meeting of the National Council on Measurement in Education, Seattle.

Robey, R.R., & Barcikowski, R.S. (1992). Type I error and the number of iterations in Monte Carlo studies of robustness. *British Journal of Mathematical and Statistical Psychology 45*, 283-288.

Schafer, J.L. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall.

Schafer, J.L. (2001). Multiple imputation FAQ page. [Online]. Available: http://www.stat.psu.edu/~jls/mifaq.html.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* (Report No. ETS-RR-94-5). Princeton, NJ: Educational Testing Service.

Stocking, M.L. (1990). Specifying optimum examinees for item parameter estimation in item response theory. *Psychometrika 55*, 461-475.

Stocking, M.L. (1988). *Scale drift in on-line calibration*. (Report No. 88-28-ONR). Princeton, NJ: Educational Testing Service.

Sympson, J.B. & Hetter, R.D. (1985). Controlling item exposure rates in computerized adaptive testing. *Proceedings of the 27th annual meeting of the Military Testing Association* (pp. 973-977). San Diego, CA: Navy Personnel Research and Development Center.

Thomas, N. & Gan, N. (1997, Winter). Generating multiple imputations for matrix sampling data analyzed with item response models. *Journal of Educational and Behavioral Statistics 22*(4), 425-445.

Wainer, H. & Mislevy, R.J. (1990). Item response theory, item calibration, and proficiency estimation. In H. Wainer, *Computerized adaptive testing: A primer*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Way, W. D. (1998). Protecting the integrity of computerized testing item pools. *Educational Measurement: Issues and Practice 17*, 17-27.

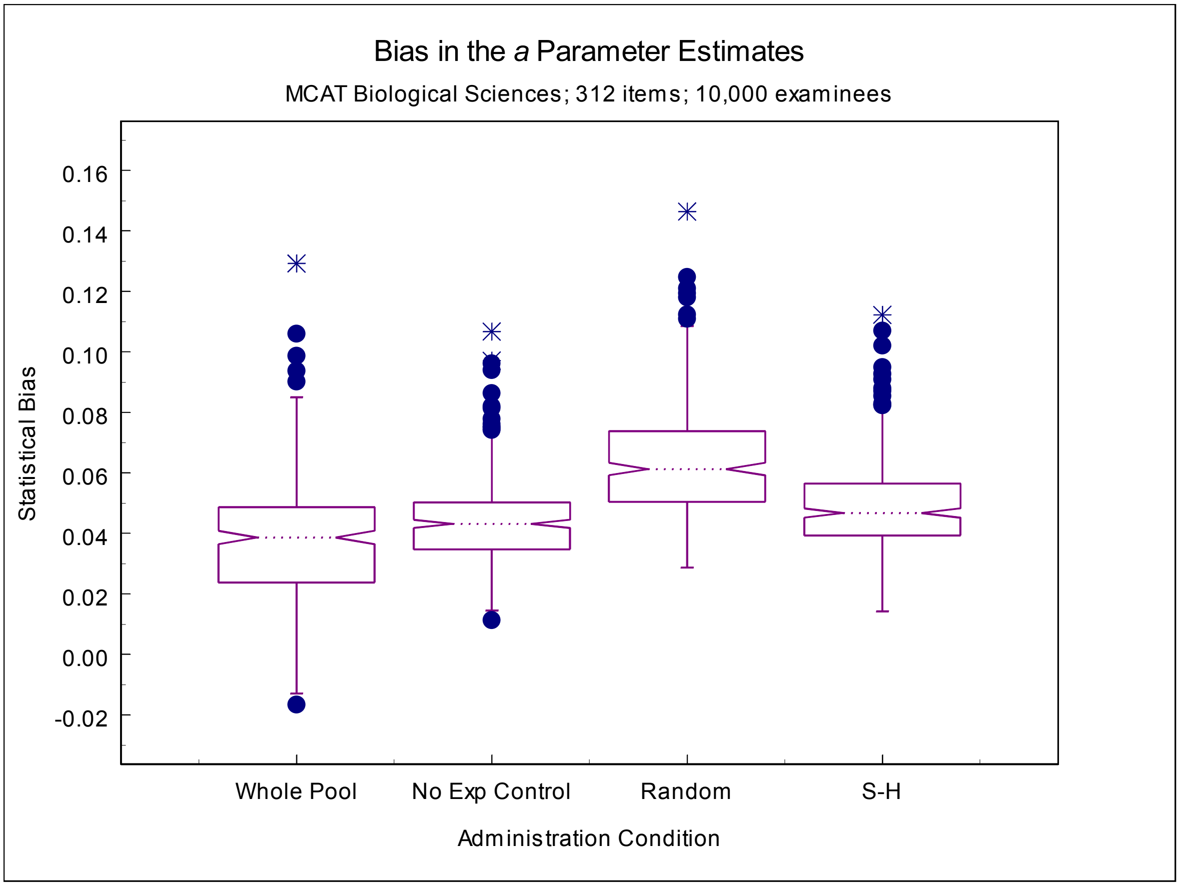

Figure1

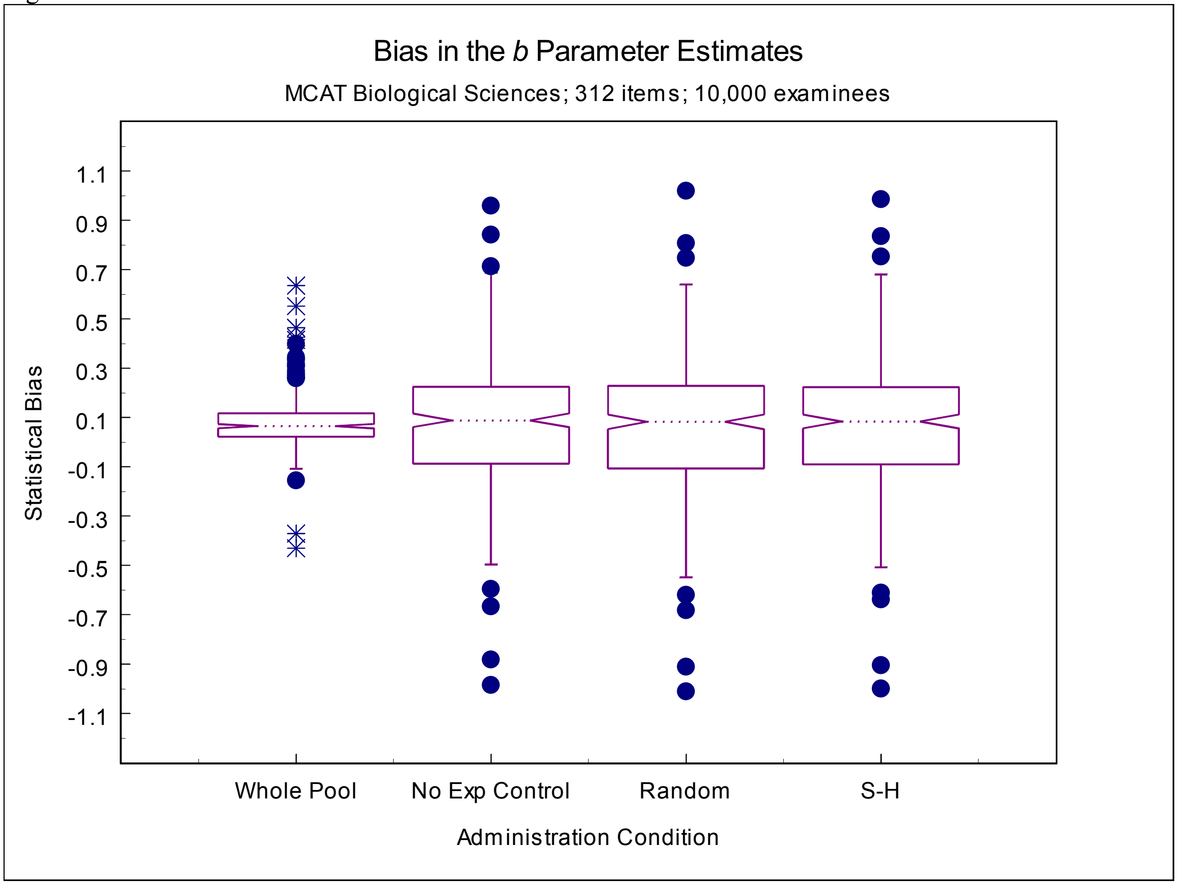

Figure 2

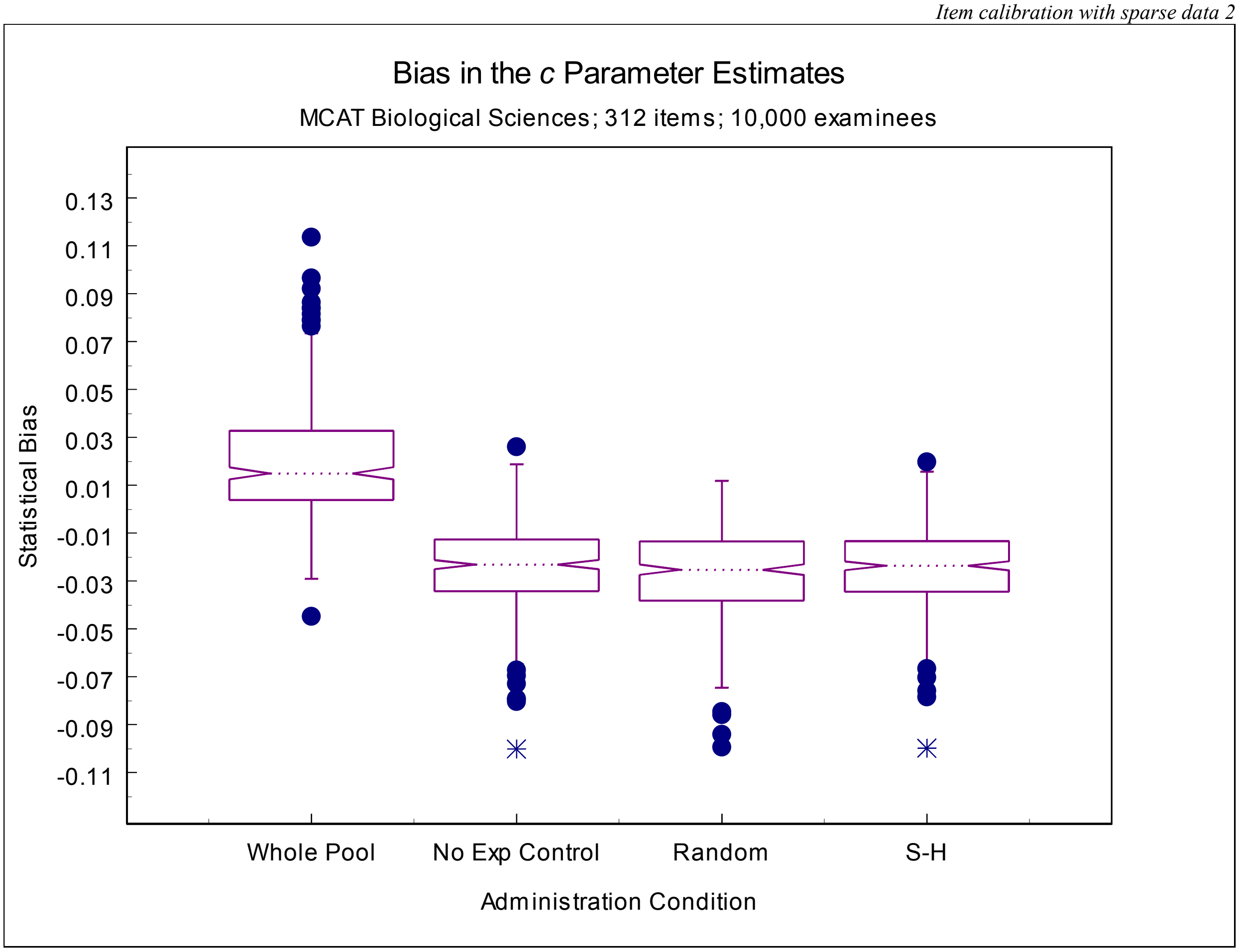

Figure 3

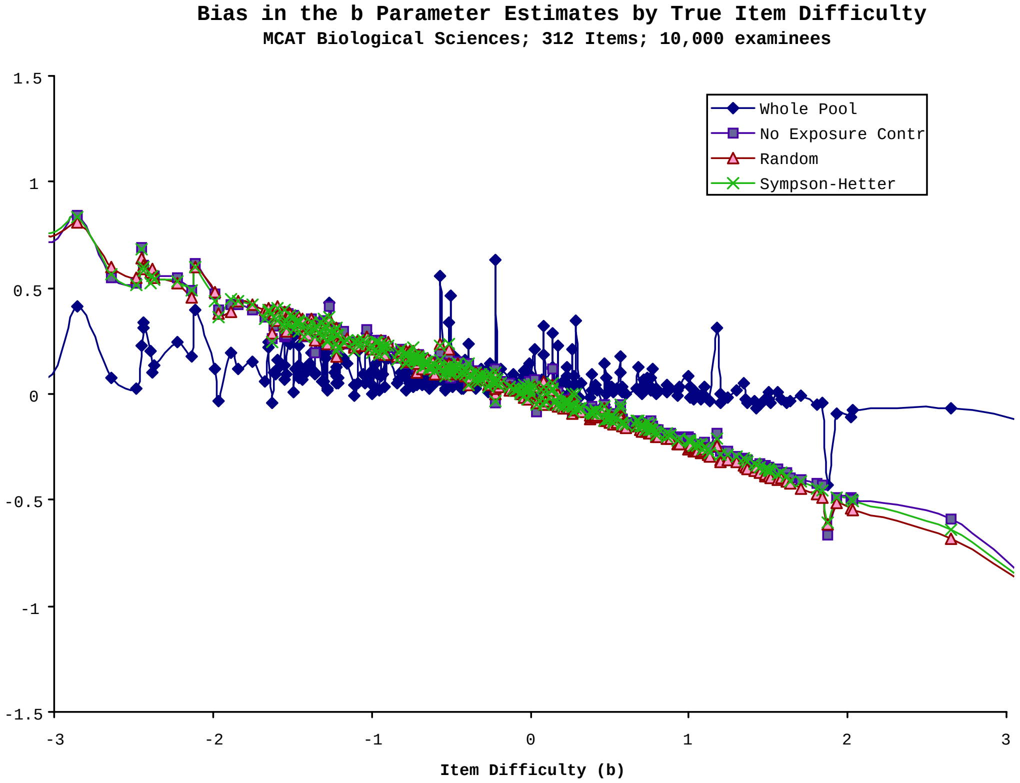

Figure 4

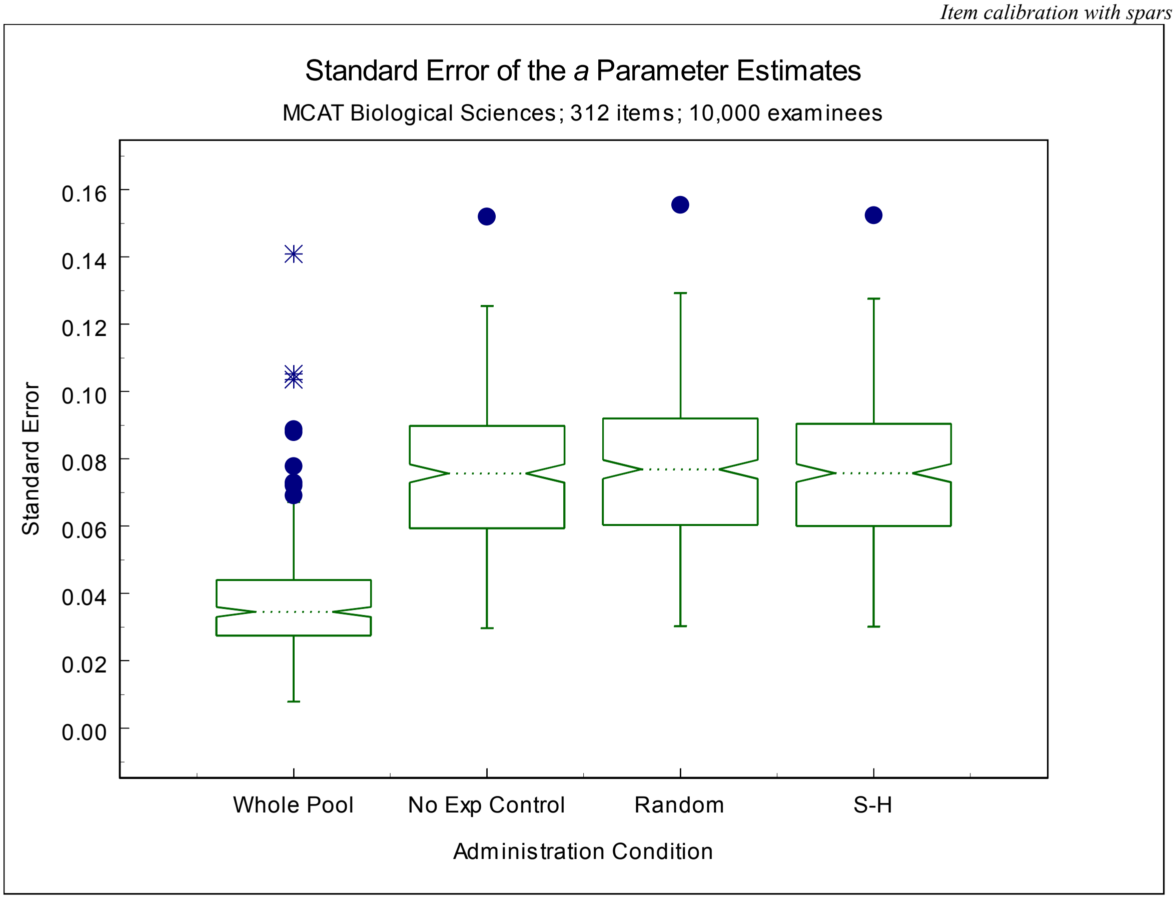

Figure 5

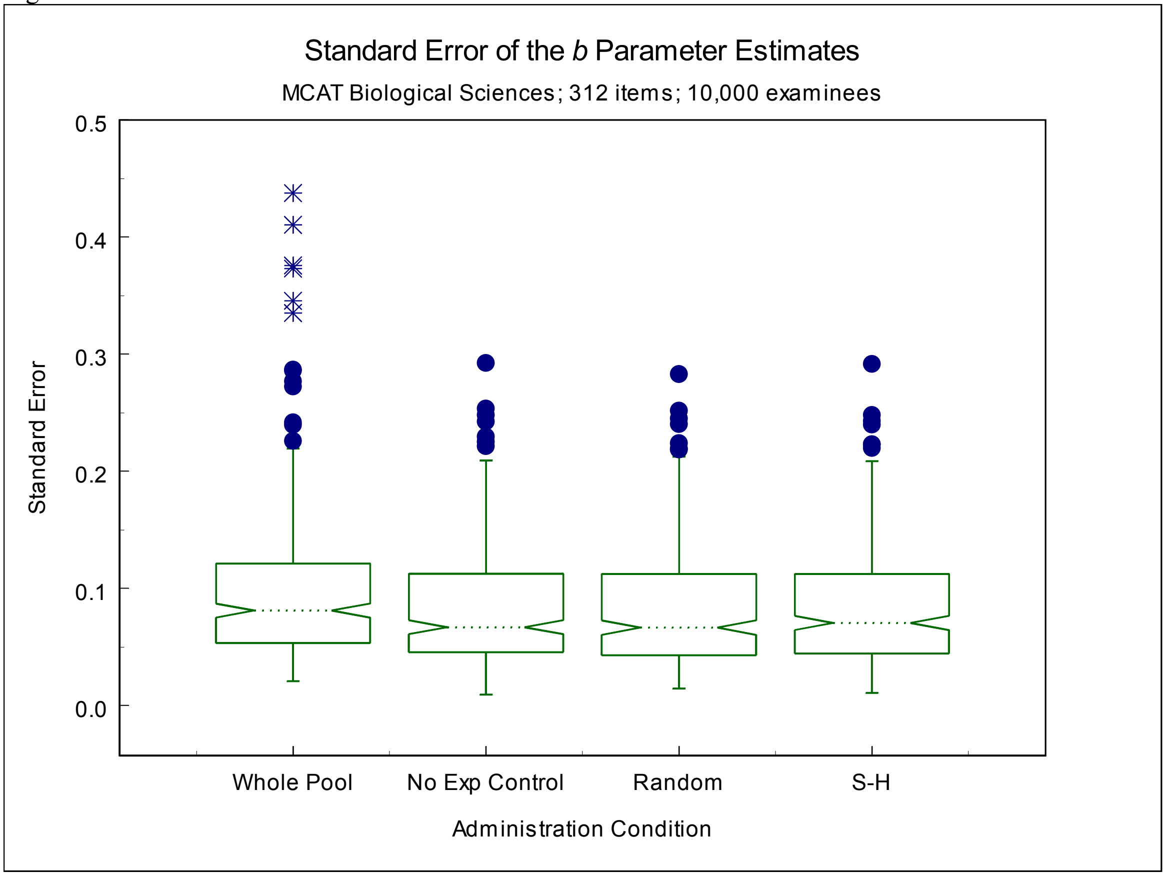

Figure 6

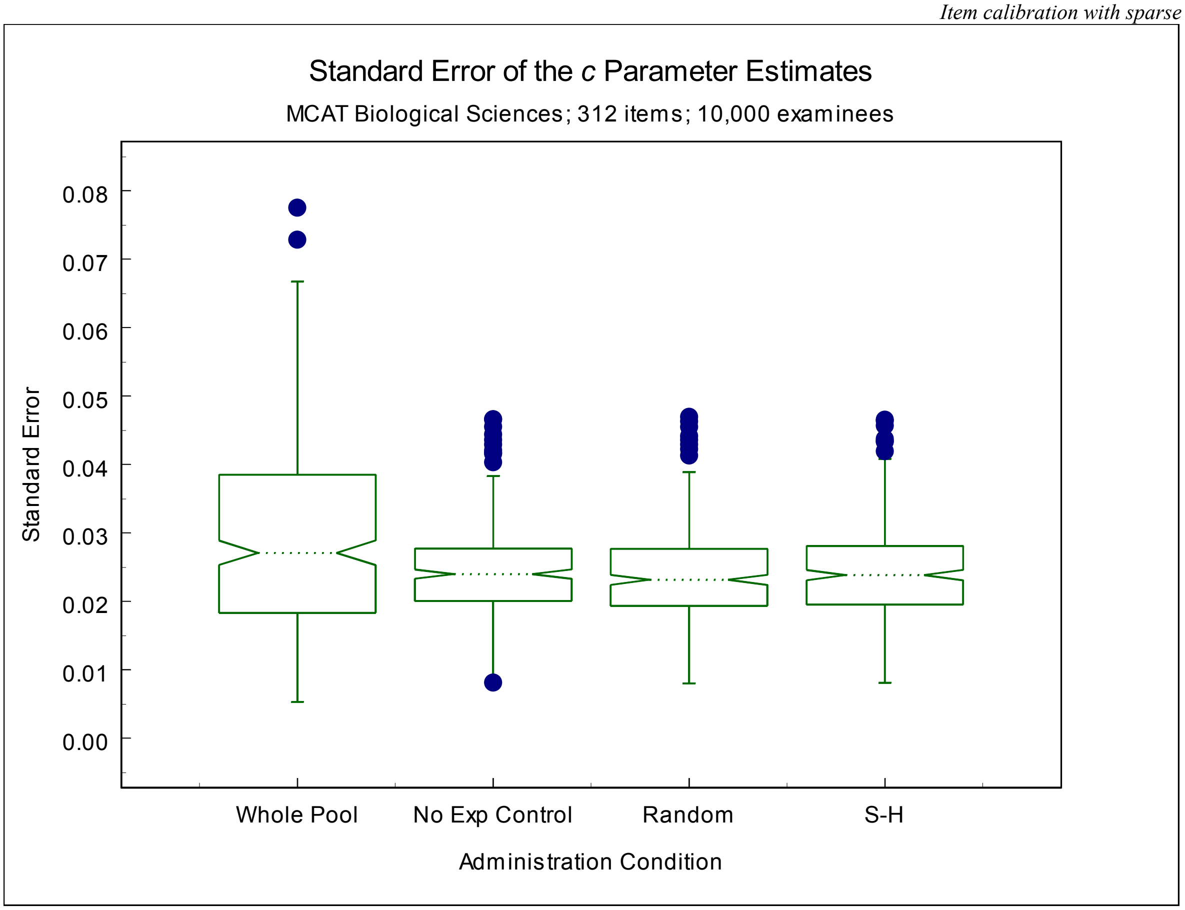

Figure 7

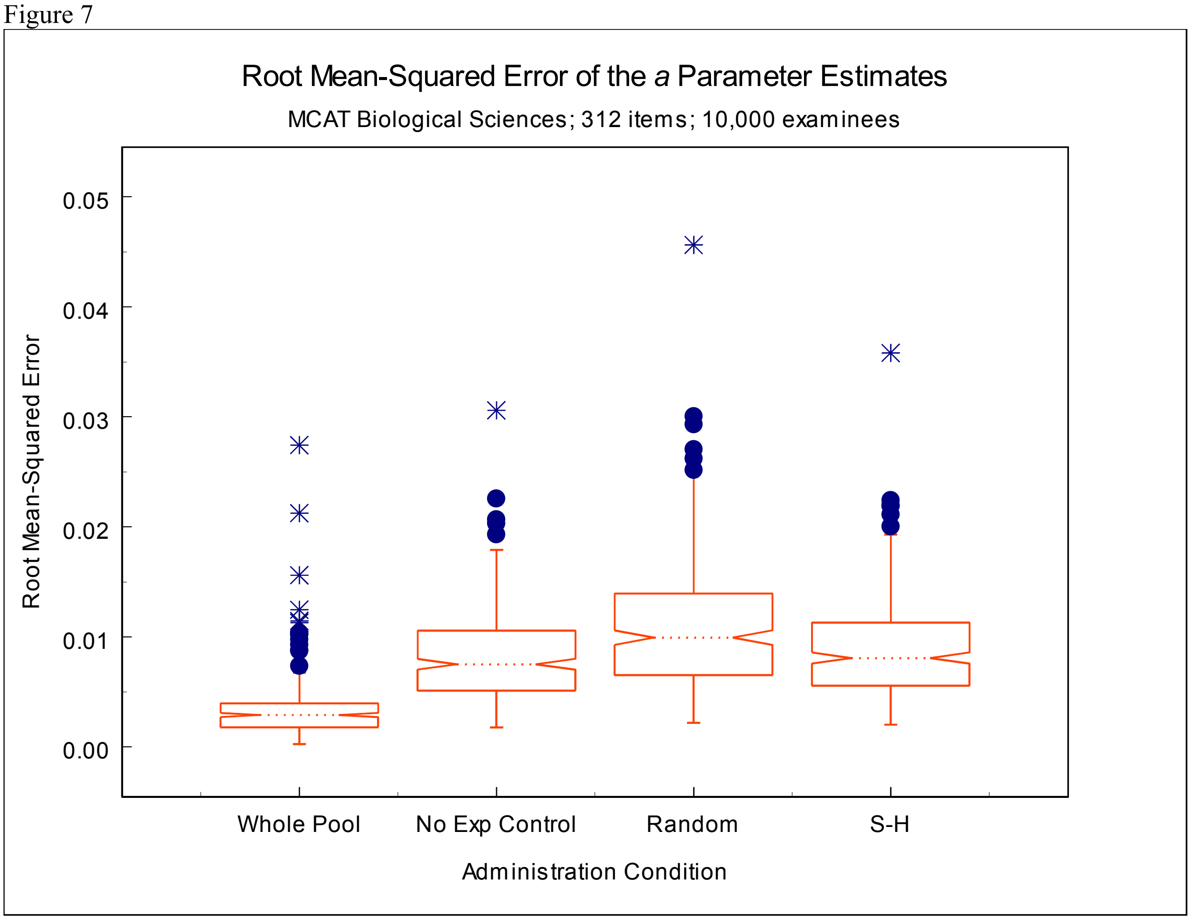

Figure 8

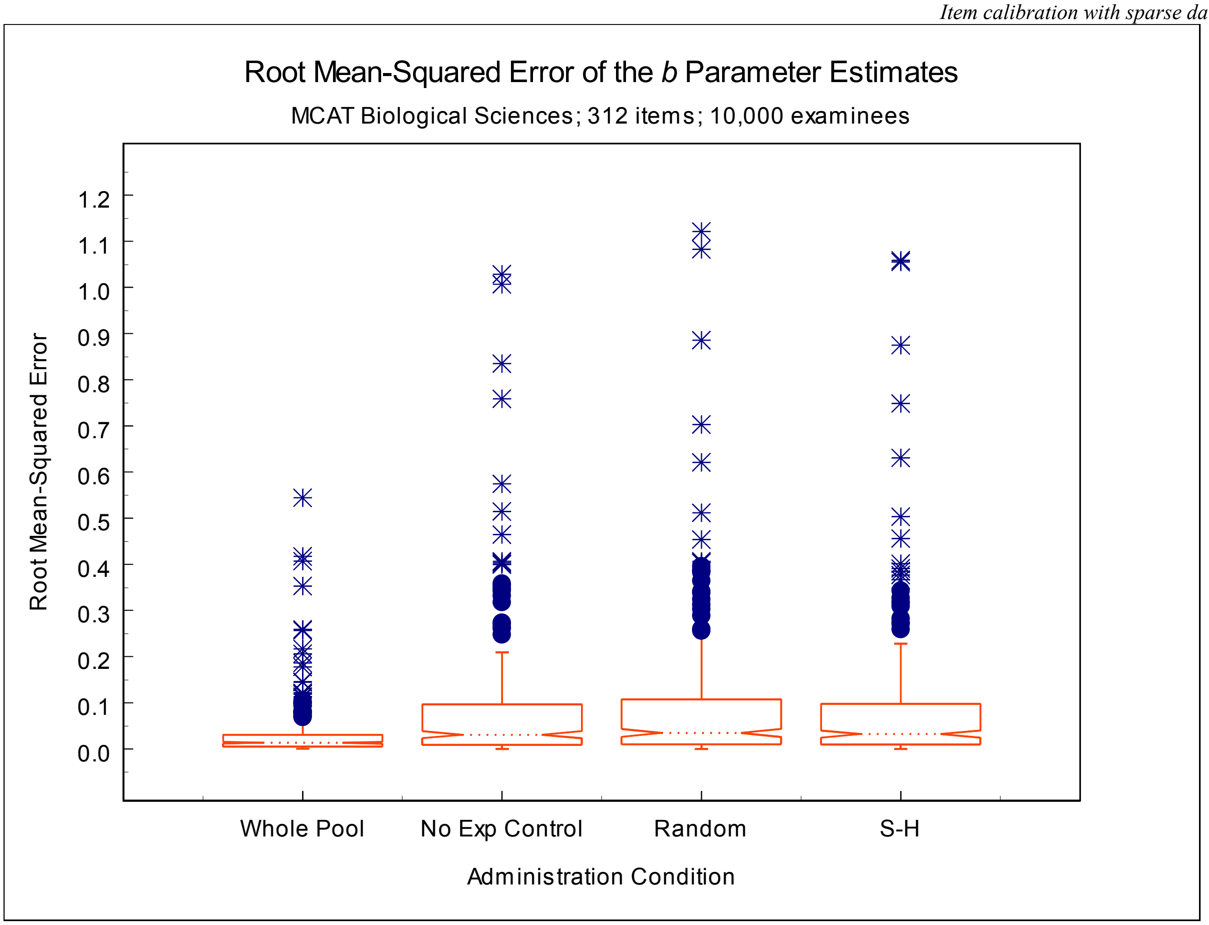

Figure 9

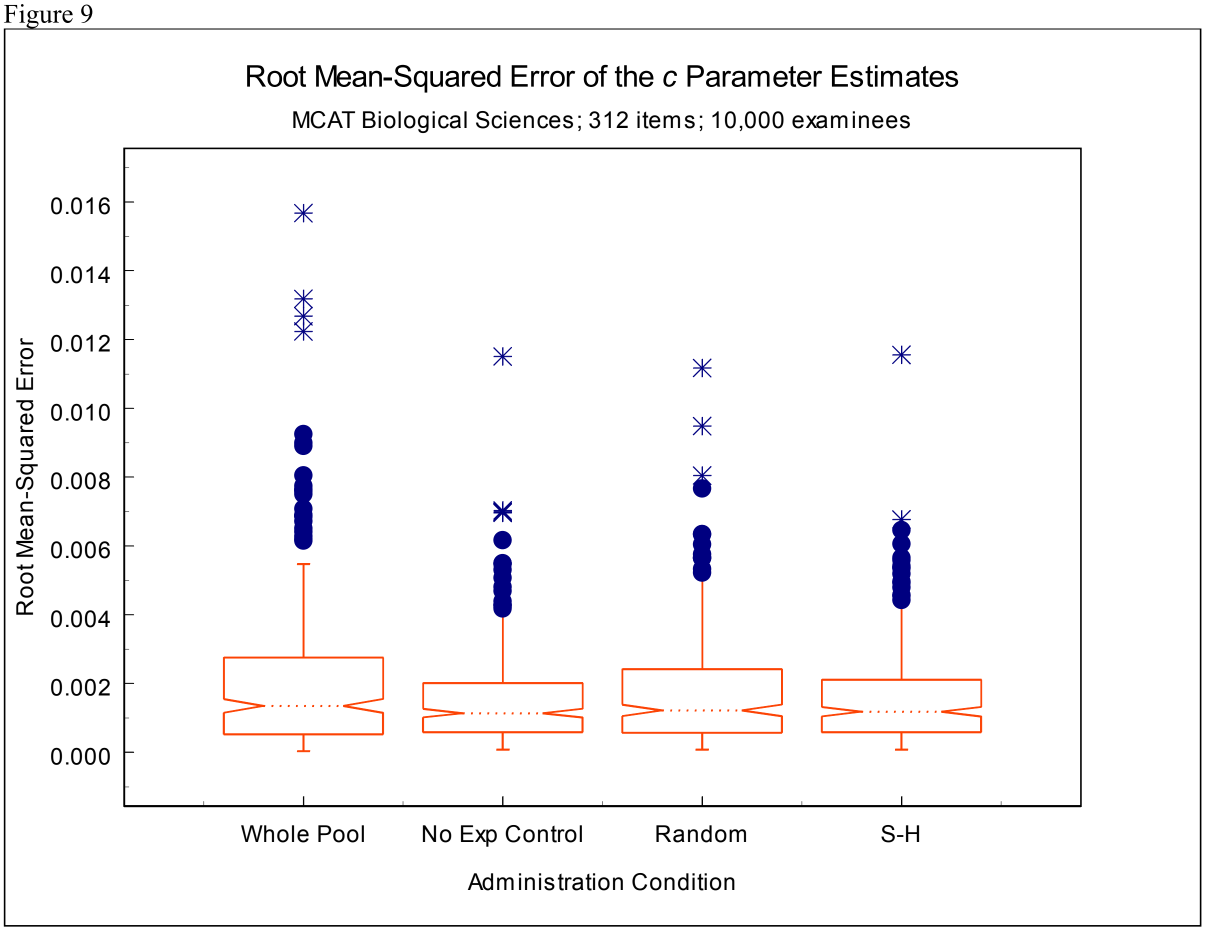

Figure 10