RECALIBRATION OF IRT ITEM PARAMETERS IN A CAT:
SPARSE DATA MATRICES AND MISSING DATA TREATMENTS

J. Christine Harmes
Cynthia G. Parshall
Jeffrey D. Kromrey

University of South Florida

Paper presented at the annual meeting of the National Council on Measurement in Education, April 22-24, 2003, Chicago.

BACKGROUND AND RATIONALE

Item parameter accuracy is critical in a computerized adaptive test (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 the final ability estimates. If these parameters are inaccurate or unstable, the integrity of the computerized testing program is in jeopardy. Calibration of item parameters based on examinee response data from operational CAT administrations has proven problematic. Due to the adaptive testing algorithm, the resultant data matrix is sparse, and could lead to parameter estimation inaccuracies.

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 on computer is best (Haynie & Way, 1995; Ito & Sykes, 1994). Problems that may affect parameter estimation 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). A testing program could elect to begin operational CAT administration using paper-and-pencil calibrations and then recalibrate when sufficient data have been collected. Recalibration is also recommended periodically in order to address possible scale drift (Stocking, 1988).

The major sources of difficulty in recalibrating parameters for a test that is being administered as a CAT lie in the restriction of the 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 – perhaps as many as 12 times the average test length (Stocking,1994).

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 will 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). In order to recalibrate items using data from a CAT administration and achieve comparable accuracy to those calibrated from 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. Strategies that might be useful include *maximum likelihood (ML) estimation via the EM algorithm* and *multiple imputation*. For nonignorable missing data, the EM algorithm is used in an iterative process for calculating ML estimates. The incorporation of the missing data mechanism (i.e., the process that leads to missingness in the data matrix) into the estimation algorithm holds promise for substantially reducing statistical bias. 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).

PURPOSE

In order to ensure a smooth transition of standardized testing programs to CAT administration and to allow for item recalibration in existing CAT programs, issues related to the quality of the item parameter estimates should be addressed. The purpose of this study was to investigate the relative effectiveness of missing data treatments applied to the sparse data matrices obtained from CAT administrations. The effectiveness of treating the missing data prior to item calibration was investigated under a variety of conditions of test length, sample size and item selection algorithms.

DATA SOURCE

Two datasets were used in this study. For the first dataset, actual examinee response data were obtained from the Medical College Admissions Test (MCAT) Biological Sciences test. The data consisted of results from six forms of this paper-and-pencil test with 3,000 to 6,000 examinee responses for each item. This yielded a dataset of 312 items from the content areas of biology and organic chemistry. These items were administered as passage-based items. The study was replicated with data from the ACT Mathematics test. These data come from 8 forms of a 60-item multiple-choice test that was administered in paper-and-pencil format. The resulting item pool consisted of 480 discrete items, representing six content areas in mathematics.

METHOD

The study can be conceptualized as having occurred in two phases. The first phase involved calibration of the baseline item parameter estimates, simulation of computerized test administrations, and analysis of the resulting data matrices. The second phase focused on missing data treatments. The data matrices obtained at the end of the first phase were “filled in” or estimated using the two missing data treatments. Item parameter estimates were calculated based on each approach, and the effectiveness of the missing data treatments was evaluated.

Baseline Item Parameter Estimates

For the MCAT dataset, phase one began with calibration of the item parameters based on existing data from paper-and-pencil administration of the tests. Existing sets of scored data from the MCAT tests were presented to BILOG (Mislevy & Bock, 1990) for IRT item parameter calibration. Once the items from the various test forms were calibrated based on paper-and-pencil test administration data, the data files were then combined to construct an item pool. Because the six test forms were administered to randomly equivalent groups, and were all calibrated using BILOG, the items from each test form could then be regarded as being on the same scale, and could thus be combined into a single item pool (Mislevy & Bock, 1990).

A similar process had previously been conducted on eight ACT Mathematics test forms. This yielded a pool of item parameter estimates for 480 items. These parameter estimates were obtained from ACT, Inc. and had been used in previous simulation studies (e.g., Yi, 1998).

For both datasets, these item parameter estimates served as baseline data to which the item parameter estimates calibrated later in this study were compared. The use of parameters calibrated from actual examinee responses on paper-and-pencil tests to conduct adaptive test simulations follows recommendations from and procedures used by other researchers (e.g., Ban et al., 2001; Davey et al., 1997; Ito & Sykes, 1994).

Test Administration Simulation

Once the item pools had been prepared, the next step was to simulate test administrations. Adaptive test administrations were simulated under four conditions, three of which resulted in sparse response data matrices. Item administration conditions that were simulated included: whole pool, random item selection, CAT with no exposure control, and CAT with Sympson-Hetter exposure control. Whole pool administration served as a 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 data that were *missing completely at random* (MCAR; missingness is unrelated to the variables being studied), and served as another baseline comparison. The two types of CAT administration (no exposure control and Sympson-Hetter exposure control) were simulated as the conditions under investigation in which the data were *missing at random* (MAR; missingness is related to the observed data). Within all four administration conditions, the number of examinees was manipulated. In the random and the two CAT conditions, proportional test length was also manipulated. 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, three proportional test lengths were investigated: (a) 1/6 of the item pool, (b) 1/9 of the item pool, and (c) 1/12 of the item pool. The minimum number of examinees needed for 3-PL item parameter calibration is approximately 1, 000 (Wainer & Mislevy, 1990) thus this number was chosen as the first level. Hsu, Thompson, and Chen (1998) recommend that approximately 10 times the number of examinees are needed to calibrate with online data in order to have comparable accuracy to calibration with a full dataset from paper-and-pencil administration. Based upon this recommendation, 10,000 was chosen as the upper level. The middle level, 5,000, was chosen as an approximate midpoint between the other two. In order to minimize the potential effects of sampling error, 100 replications were conducted (Robey & Barcikowski, 1992).

Examinee abilities were generated to represent a normal distribution, as is common in similar psychometric studies (e.g., Ban, et al., 2001; Davey, Nering, & Thompson, 1997; Harwell, et al., 1996; Pommerich, 2002). This represented a situation reasonably 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.

These sparse response data matrices were analyzed in terms of the degree of sparseness resulting from the various administration conditions. Before applying the missing data treatment, calibration of the sparse response data matrices was attempted using BILOG (Mislevy & Bock, 1990) in order to evaluate the severity of the sparseness problem, and to determine whether sparse datasets could be calibrated without being adjusted.

Missing Data Treatments

Drawing from the applied statistics literature on solutions for dealing with missing data, the multiple imputation technique and the EM algorithm were applied to the sparse data matrices. Ten sets of imputations were completed within each replication for the MI approach (Rubin, 1987; Schafer, 2001).

The first step in phase two was to recalibrate the items based on the simulated data. Scored responses from the whole pool administration were used to calibrate items for this control condition. To calibrate the random and CAT administration data, two missing data treatments were used: ML with EM and multiple imputation. Each technique was applied to the resulting data matrices from the random and CAT administrations. For example, from the condition in which an 80-item test was administered to 1000 examinees, the resulting data matrix was “filled in” using multiple imputation, and then parameter estimates were calibrated from this full data matrix. Under this same condition, ML estimates of the item parameters were calculated using the EM algorithm, taking into account the missing data points. These two sets of parameters, obtained from different treatments, were subsequently compared to the true parameters for accuracy and stability.

Multiple Imputation

For multiple imputation, 10 sets of imputations were used to create 10 complete item response data matrices within each replication, for each condition. In the case of CAT, the missingness in the dataset is due to the examinee’s ability estimate (θ) and the missing response values were dichotomous, (0,1) indicating either a correct or incorrect response to each item. Thus, the scored examinee responses (0,1) were imputed for all items that were not administered in the CAT. To conduct the multiple imputations, the final estimate of examinee ability and the variance of that estimate (obtained at the conclusion of the simulated CAT) were used to create a posterior distribution of ability for each examinee. From this posterior distribution, 10 values of θ were randomly selected. Each of these 10 ability values was then used to probabilistically impute responses to the items that were not administered to that particular examinee. The 10 complete response matrices were subsequently calibrated using BILOG (Mislevy & Bock, 1990), and the results were combined following procedures recommended by Schafer (2001). The average of each parameter value across the 10 calibrations from the imputed datasets was used as the final item parameter estimate for the MI approach.

ML Estimation with the EM Algorithm

Sparse response data matrices from the random and two CAT administration conditions were first presented to a FORTRAN program that performed the expectation phase of the EM algorithm. This program probabilistically filled in the sparse response data matrix with imputed values, similar to a single imputation technique. In the maximization phase, this full data matrix was then presented to BILOG for item parameter estimation. The item parameter estimates returned from BILOG were then sent back to the expectation program to fill in the sparse data matrix with a new set of imputed values. This process was carried out iteratively until a convergence criterion was met.

Once the calibrations were complete, the final item parameter estimates from each administration condition were evaluated in terms of accuracy and stability.

RESULTS AND CONCLUSIONS

Item parameter estimates calculated from the results of the study conditions were compared to the "true" item parameters calibrated from the original paper-and-pencil data. Results were examined in terms of the sparseness of the response data matrices, and the bias and standard errors of the item parameter estimates. Results will be presented for only the largest proportional test length (1/6) and the largest sample size (10,000) for both datasets. Complete results are available from the first author.

Sparseness

Prior to attempting item calibration, the initial, sparse response data matrices from the random, CAT with no exposure control, and CAT with Sympson-Hetter exposure control were analyzed. The first step was to examine the nature of the sparseness. Figures 1 and 2 illustrate the percentage of missing data for each item in the pool, by true item difficulty (*b* value). The horizontal dotted line represents the average percentage of missingness for the pool, given the proportional test length (e.g., for a 40-item test from a 480 item pool, the average percent of missingness would be 92%).

The conditions studied represent an average range of missingness from 83 to 92 percent. In the MCAT data, the percent of missing data for an individual item ranged from approximately 70 to 100. Across the various administration conditions, the amount of missing data was distributed evenly across true difficulty. The range of missingness was much greater for the ACT data, with missingness from zero to 100 percent. In this dataset, missingness varied across administration conditions. For the random administration and Sympson-Hetter exposure control methods, sparseness was similar across the difficulty range. However, for the no exposure control method (i.e., maximum information item selection), sparseness was much less near the middle of the difficulty range. That is, very easy and very difficult items were not administered frequently, whereas items of medium difficulty were administered much more often. The constraints of the passage-based item selection algorithm probably contributed to the fact that the no exposure control condition did not show a similar pattern in the MCAT data.

Test Length Variations

Variations in test length resulted in different patterns of sparseness. Longer proportional test lengths resulted in more items that had a smaller percent of missing data. These items for which sparseness decreased, tended to be near the middle of the difficulty range. This was especially true under no exposure control administration in both datasets, and was most noticeable in the ACT pool. Sparseness was spread more evenly across the difficulty range under random administration, followed by Sympson-Hetter exposure control.

Numbers of Examinee Variations

Variations in number of examinees in the sample had very little effect on the average levels of sparseness in the response data matrices. Larger numbers of examinees resulted in more items with slightly less sparseness.

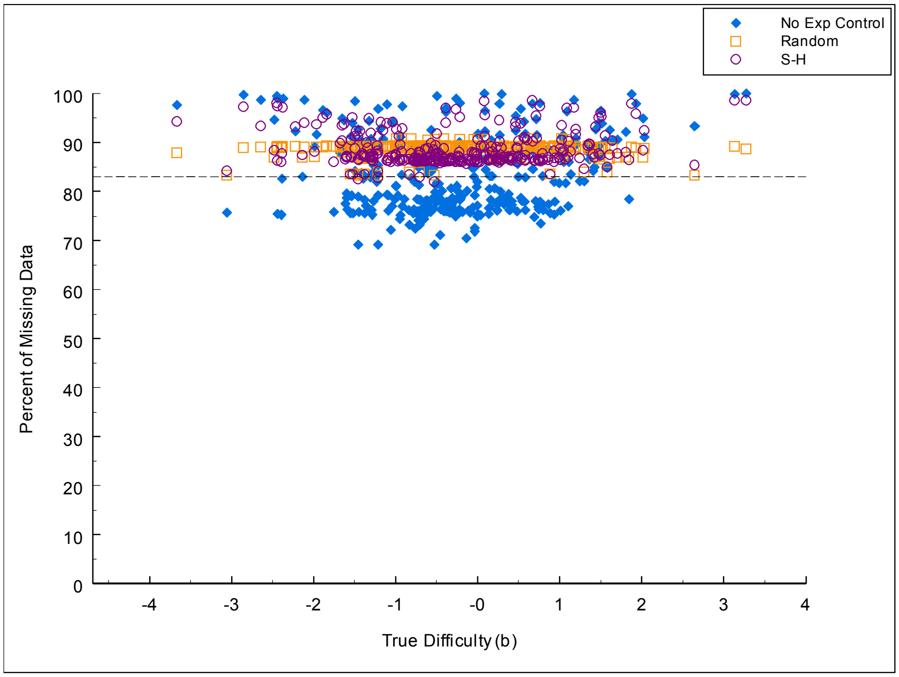

Figure 1. Sparseness in the response data matrix by true item difficulty; MCAT Biological Sciences; 1/6 of 312 items; 10,000 examinees.

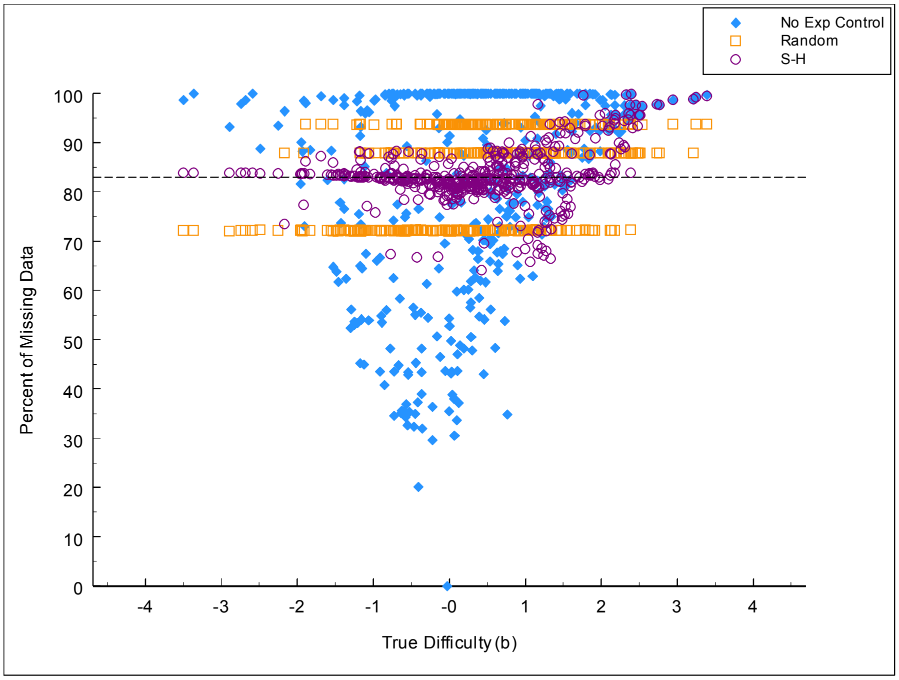

Figure 2. Sparseness in the response data matrix by true item difficulty; ACT Mathematics; 1/6 of 480 items; 10,000 examinees.

Attempted Calibration of Sparse Response Data Matrices

Before applying missing data treatments to the data matrices, item calibration was attempted for each sparse dataset. The random, CAT with no exposure control, and CAT with Sympson-Hetter exposure control conditions all resulted in sparse response data matrices. For each replication, these sparse datasets were presented to BILOG for calibration. Calibration failed in all cases. With most datasets, BILOG completed Phase One, but failed to complete Phase Two, and thus no item parameters were output. In a few cases, Phase Two was completed, however the *largest change* value reported by BILOG increased after each cycle, and unreasonable parameter estimates (or simply asterisks) resulted. This consistent pattern of results is a clear indication that a remedy is needed for recalibration of CAT item parameters. Since all items in the pool are involved in the calibration process, the items with a larger degree of sparseness would have had an effect on the calibration of the entire pool. That is, there may have been individual items for which BILOG would have had sufficient data for calibration. In order to calibrate those items, a subset of the items with sufficient data would have had to be sent separately to BILOG for calibration. However, this solution has drawbacks; calibration of only the subset of items with sufficient data does not solve the problem of calibrating CAT datasets. In the datasets in this study, the number of items with sufficient data for calibration would have been quite small. Obtaining new calibrations for a very small subset of the pool would create difficulties for pool usage and item exposure. Also, such a subset may not be representative of the full domain denoted by the entire item pool.

Variability Across Imputations

After multiple imputations were made in order to fill in the sparse data matrices (under the multiple imputation condition), the resulting full datasets were calibrated using BILOG. Within each replication, variance across the item parameter estimates resulting from each set of imputations was analyzed. 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.

There is virtually no difference in the average total variance of the *a* parameter across the three CAT administration conditions within a given level of test length and number of examinees, for the MCAT Biological Sciences. The average total variance for the *a* parameter decreased as sample size (number of examinees) increased, however differences in test length across the same sample size did not affect the total variance across imputations. The highest maximum total variance for the *a* parameter (0.5723) was seen under random administration with a proportional test length of 1/6 of the pool with 1,000 examinees. For the *b* parameter, the highest average value of total variance across imputations (0.0739) was found under no exposure control administration with 1,000 examinees and a proportional test length of 1/6 of the pool. The smallest value for total variance (0.0161) occurred under random administration with 10,000 examinees and a proportional test length of 1/12. Similarly, with the *c* parameter, the smallest average value for total variance (0.0018) was seen under random administration with 10,000 examinees (1/12 of the pool). The highest average value (0.0062) occurred under no exposure control administration with 1/6 of the pool and 1,000 examinees.

For the ACT Mathematics Test, the smallest amount of total variance in the *a* parameter across imputations (0.0057) was seen under both no exposure control and Sympson-Hetter exposure control with 10,000 examinees and a proportional test length of 1/6 of the item pool. The smallest value of total variance for the *b* parameter (0.0080) occurred under random administration with 10,000 examinees and a proportional test length of 1/12. Similarly, the smallest value for total variance in the *c* parameter (0.0008) was found under random administration with 10,000 examinees and both 1/6 and 1/12 of the pool. Largest average values for total variance across all three parameters were observed under conditions with 1,000 examinees. Under these conditions, extreme values were found in this dataset for maximum total variance of the *a* and *b* parameters, most notably under proportional test lengths of 1/6 and 1/9. This appears to be a result of filled-in datasets that failed to converge in the item calibration process. Although this did also occur with the MCAT pool under the same conditions, it was much less common and was easily corrected. This does, however present a need for caution when attempting to use this relatively small sample size. In the ACT data, this problem carried through as item parameter estimates were averaged and compared to baseline parameters.

Bias

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 and EM estimation.

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. The calibration of results based upon the whole pool is included as a reference distribution.

For each item parameter, two types of graphs are used to present results. First are notched box-and-whisker plots, used to show the distributions of the respective statistic (e.g., the distribution of item parameter estimate bias or standard error across all items in the pool) under each administration condition. The box distinguishes the 25th through 75th percentiles of the distribution, or the interquartile range. The horizontal dotted line in the middle of each box denotes the median of the distribution, while the notches represent the magnitude of the standard error around this median. The vertical lines extending from the box represent the spread of the top 25% and bottom 25% of the distributions. Circles at the ends of these vertical lines indicate that there were observations more than 1.5 inter-quartile ranges above or below the box, while asterisks indicate observations more than 3 inter-quartile ranges above or below. In the second type of graphs, given for bias, the value of bias is plotted for each item, by that item's true parameter value (*a*, *b*, or *c*) as specified.

Item Discrimination (a)

MI Estimation

MCAT Biological Science. The random condition (1,000 examinees and 1/12 of the pool) yielded the highest average amount of statistical bias for the *a* parameter (0.2429), while the whole pool condition with 10,000 had the lowest average bias (0.0367). This held true across all testing conditions. The highest maximum bias was found in the random condition (0.4003) with 1.000 examinees and a proportional test length of 1/12 of the pool. Overall bias in the *a* parameter was positive, that is, items were estimated as being more discriminating than they actually were.

For sample sizes of 1,000 examinees, increasing test length had a minimal effect on bias in the *a* parameter. For sample size of 5,000, increasing test length reduced the overall amount and variability of the average bias. For example, under no exposure control, average bias was 0.1619 for a proportional test length of 1/12. When the proportional test length was increased to 1/6 of the pool, average bias decreased to 0.1058. Similarly, with 10,000 examinees, an increase in test length corresponded to a decrease in the level and variability of bias.

Across all levels of proportional test length, increasing the number of examinees led to a decrease in the bias of the *a* parameter estimates. Under random administration with 1/6 of the pool, average bias was 0.1855 for 1,000 examinees, and 0.1084 for 10,000 examinees.

ACT Mathematics. The whole pool with 10,000 examinees condition yielded the highest average amount of statistical bias for the *a* parameter (-0.0731), while the random condition with 1,000 examinees and 1/9 of the pool had the lowest average bias (0.0061). The highest maximum bias (0.2342) was found in the whole pool with 1,000 examinees condition. In this dataset, overall bias on the *a* parameter was negative. Thus, items were estimated as being less discriminating than they truly were.

Increasing test length led to slight changes in bias. Under no exposure control with a sample size of 5,000, increasing test length modestly increased average bias from –0.0511 to –0.0579. Similarly, with 10,000 examinees, an increase in test length corresponded to an increase in the level of bias. With random administration and 1/12 of the pool, average bias was –0.0175. Increasing the proportional test length to 1/6 increased the average bias to –0.0510. In the 5,000 and 10,000 sample sizes, increase in test length also resulted in the bias across item administration methods becoming more similar.

In general, increasing sample size led to an increase in the average bias value, along with a decrease in the variability and magnitude of maximum values. For the shortest proportional test length (1/12 of the pool) under no exposure control, an increase in the number of examinees from 5,000 to 10,000 resulted in an increase in the average bias value from –0.0511 to –0.0636. However, the maximum bias values seen under this condition decreased from 0.0379 with 5,000 examinees to 0.0136.

EM Estimation

As seen in Figures 3 and 4, bias in the *a* parameter was much greater with EM estimation than with multiple imputation. In the case of the MCAT dataset, the average EM estimates were much more positively biased, with greater variability. In the ACT dataset, the average EM estimates were much more negatively biased than the multiple imputation estimates. This difference was especially large with the ACT dataset, and the level of variability was extreme.

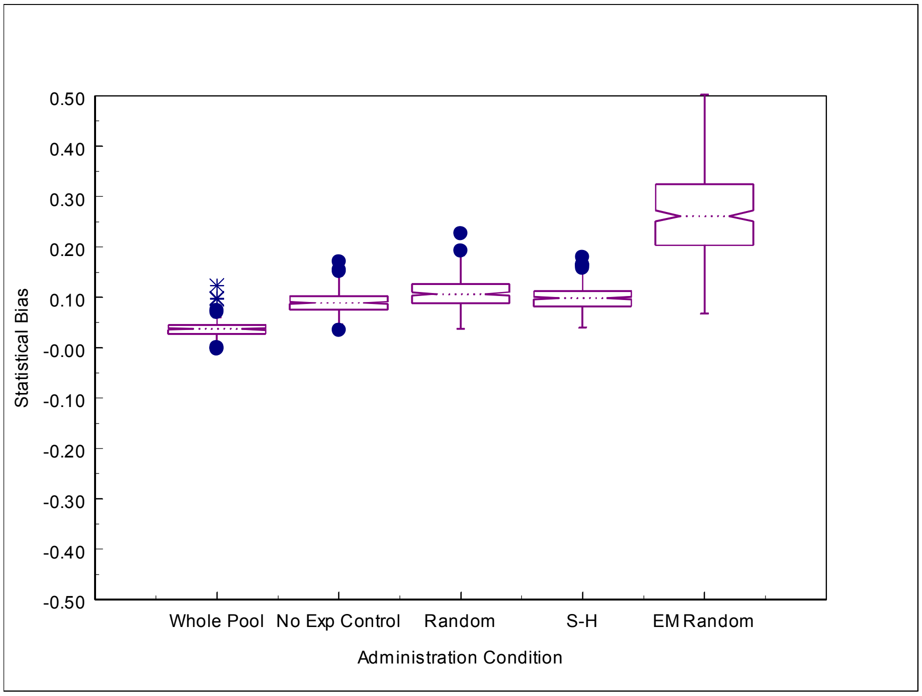

Figure 3. Bias in the *a* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

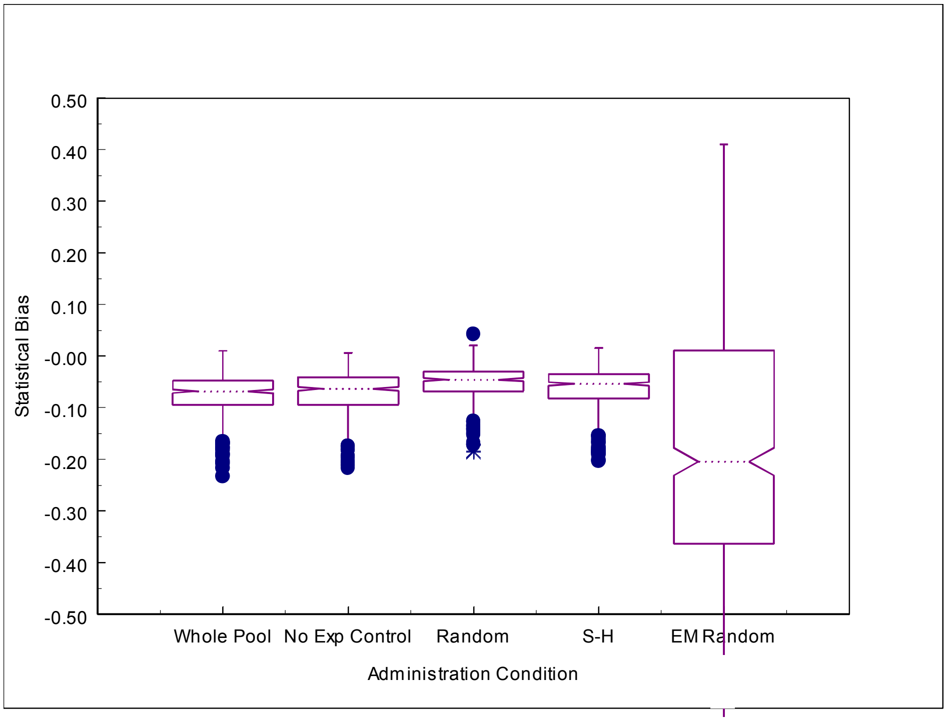

Figure 4. Bias in the *a* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Figures 5 and 6 plot the bias in *a* parameter estimates by true item discrimination. Across both datasets the conditions with longer proportional test lengths and larger sample sizes show much less bias and a reduction in variability of this bias. In the MCAT data bias in the *a* parameter increases as items are more discriminating. Conversely, for the ACT data, bias in the *a* parameter increases (negatively) as the items become more discriminating. The overall magnitude of the bias across the two datasets is fairly similar across the conditions. In the MCAT dataset the experimental conditions all result in more bias than the whole pool reference condition. With the ACT data, all of the item administration methods perform very similarly.

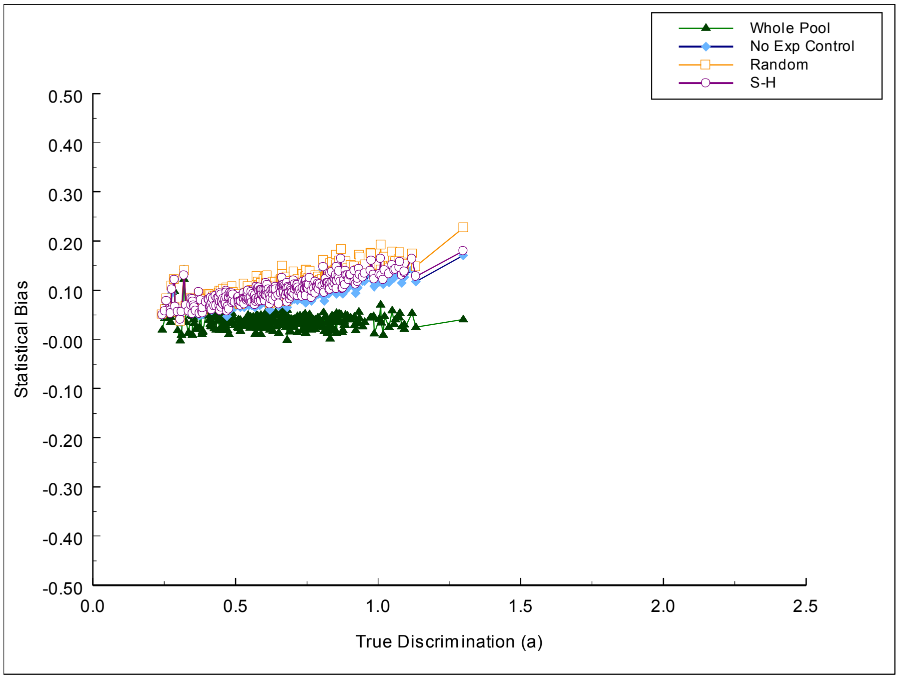

Figure 5. Bias in the *a* parameter estimates by true discrimination; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

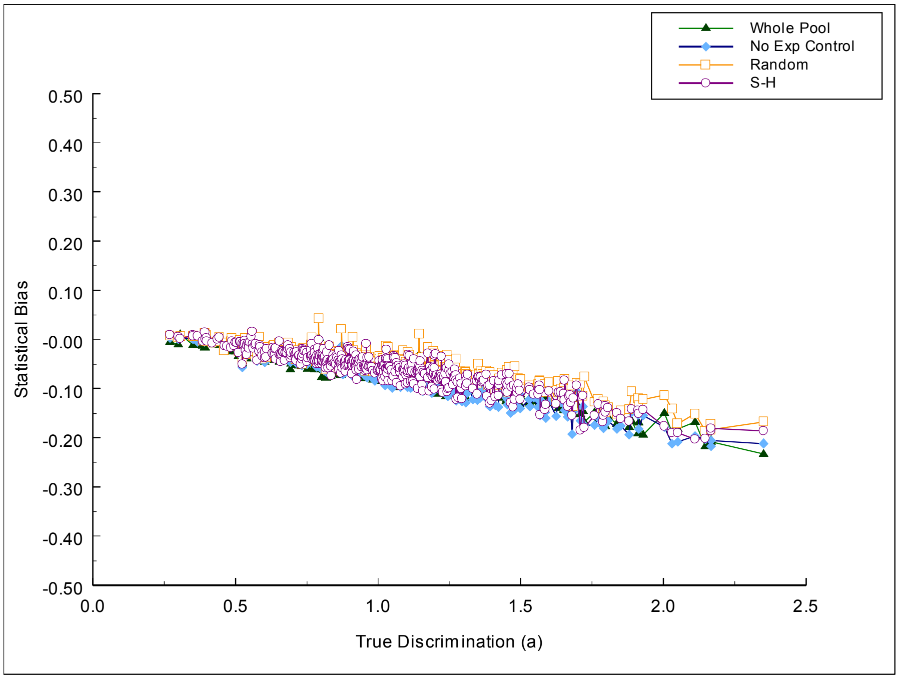

Figure 6. Bias in the *a* parameter estimates by true discrimination; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Item Difficulty (b)

MI Estimation

MCAT Biological Sciences. For the *b* parameter, average bias was positive, indicating that items were being estimated as being more difficult than they truly were. While the average values were very close to zero, items in the top and bottom 25th percentiles fell virtually equidistant from the middle of the distribution into both positive and negative bias. The largest average amount of bias occurred in the whole pool with 1,000 examinees condition (0.1268), while the random with 10,000 examinees and 1/12 of the pool condition showed the lowest average value for bias (0.0279). Bias was very similar across item administration methods.

For sample sizes of 1,000, variations in test length resulted in very little change in average bias for the *b* parameter. In the case of larger sample sizes, increasing test length resulted in a reduction in the number and level of extreme cases, while the average amount of bias changed very little. With Sympson-Hetter exposure control and 10,000 examinees, the maximum bias value decreased from 0.7873 to 0.5473 when increasing proportional test length from 1/12 to 1/6.

In the case of the shortest test length, increasing the sample size resulted in a decrease in average bias, along with a reduction in the number and severity of the outliers. Under no exposure control, average bias decreased from 0.0960 with 1,000 examinees to 0.0333 with 10,000 examinees. Similar results were seen for the middle test length. The effect of increasing number of examinees was most pronounced in the longest proportional test length, with average level and range of bias being reduced. With random administration, average bias under this condition decreased from 0.1093 to 0.0414 when increasing the number of examinees from 1,000 to 10,000.

ACT Mathematics. Overall, bias in the *b* parameter was much smaller in the ACT dataset than the MCAT dataset. Average bias was very close to zero, with a small amount of variability and no strong tendency to either over- or under-estimate item difficulty. The no exposure control condition with 1,000 examinees and 1/12 of the pool resulted in the highest average bias (0.1081), while the random condition with 5,000 examinees and 1/12 of the pool showed the lowest average value for bias (0.0025).

Within a specific sample size, changes in test length had very little effect on the average amount of bias in the *b* parameter. The only noticeable change was a reduction in the number of extreme observations with an increase in test length. For example, under Sympson-Hetter exposure control with 10,000 examinees, bias increased from 0.0211 with 1/12 of the pool to 0.0294 with 1/6 of the pool. Under random administration (10,000 examinees) the maximum value for bias decreased from 0.3387 to 0.2166 when increasing proportional test length from 1/12 to 1/6.

Increasing the number of examinees generally led to a reduction in the average bias values. With Sympson-Hetter exposure control and the middle test length, average bias decreased from 0.0866 to 0.0264 when increasing sample size from 1,000 to 10,000. A distinct change resulting from increased sample size was a decrease in the range of bias values. As the number of examinees increased, the number of outliers decreased. For example, in the case of the longest proportional test length under Sympson-Hetter exposure control, the maximum bias values decreased from 0.5469 to 0.2396 when sample size was increased from 1,000 to 10,000.

EM Estimation

As seen in Figures 7 and 8, bias in the *b* parameter was much greater with EM estimation that with multiple imputation. In the case of the MCAT dataset, the EM estimates showed slightly more average bias, and a much wider range of extreme observations. In the ACT dataset, the average EM estimates had strong positive bias, as compared to the multiple imputation estimates with virtually no bias. The variability of the EM estimates was very large, with the top and bottom 25th percentiles extending far beyond the reasonable range.

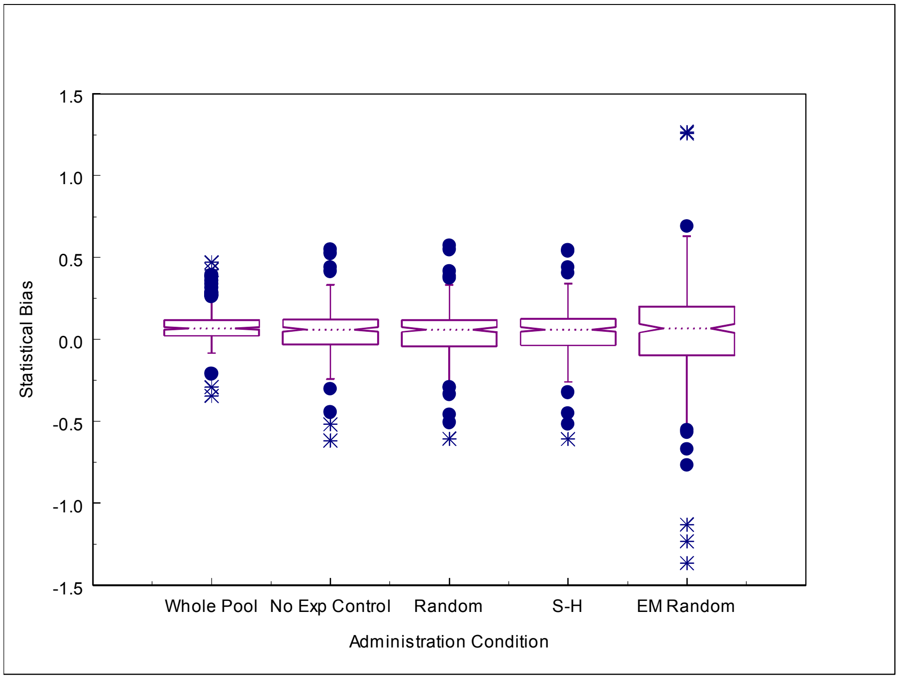

Figure 7. Bias in the *b* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

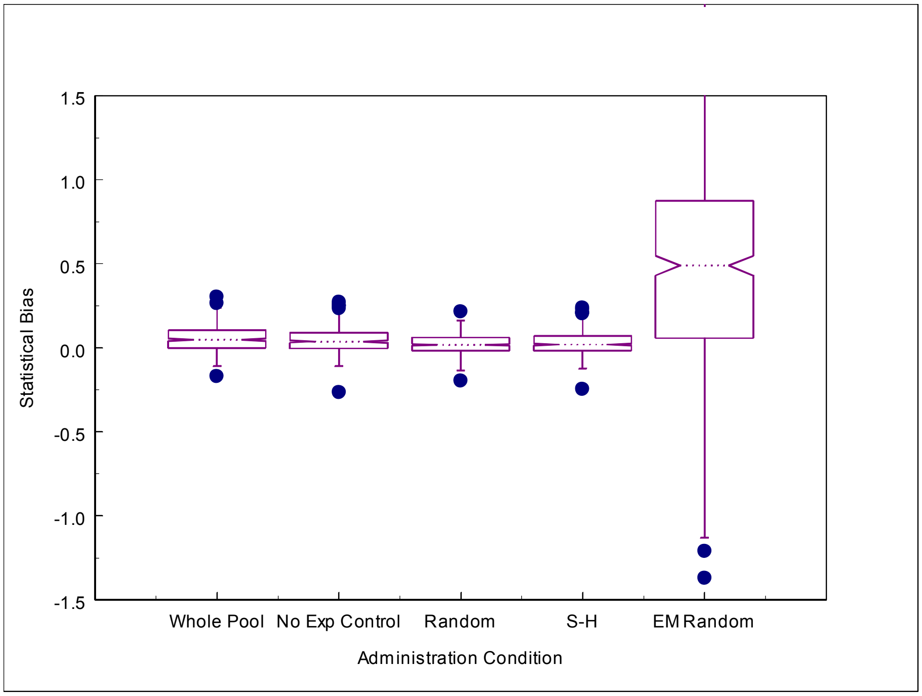

Figure 8. Bias in the *b* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Figures 9 and 10 illustrate the bias in *b* parameter estimates by true item difficulty. The magnitude of bias across the difficulty range is generally less in the ACT dataset than in the MCAT data. In both datasets, longer test lengths and greater sample sizes resulted in less bias in *b* parameter estimates. In the MCAT data positive bias occurred in less difficult items and negative bias in more difficult items. Practically, this means that easy items appeared harder and harder items appeared to be easier. This trend was similar for all item administration conditions, with the whole pool condition being less extreme than the others. The ACT data bias in the *b* parameter functioned differently. Across the conditions bias stayed very close to zero. However, as difficulty increased, there was a corresponding, although very slight, increase in positive bias.

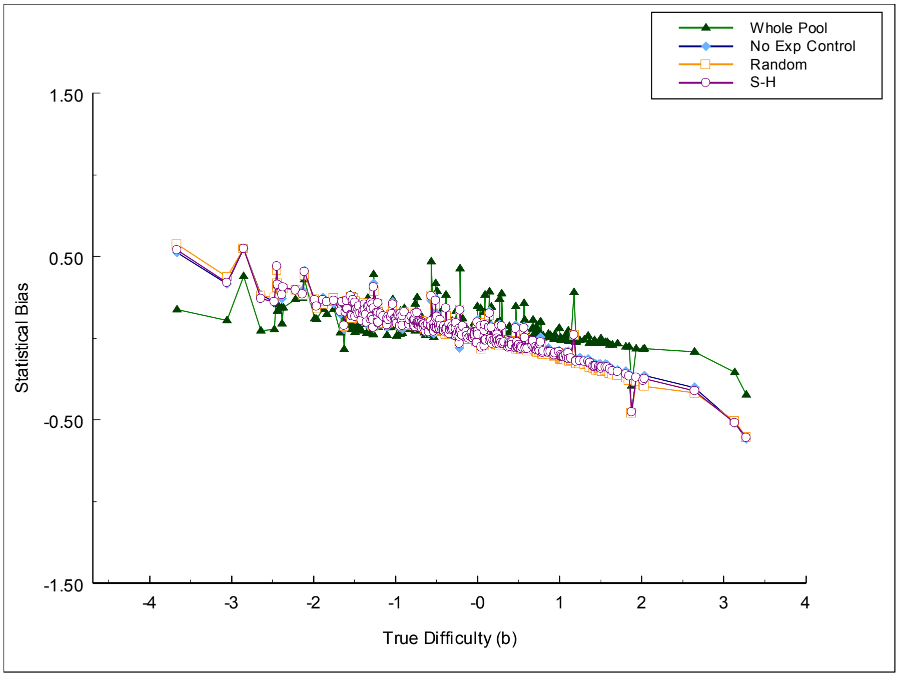

Figure 9. Bias in the *b* parameter estimates by true difficulty; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

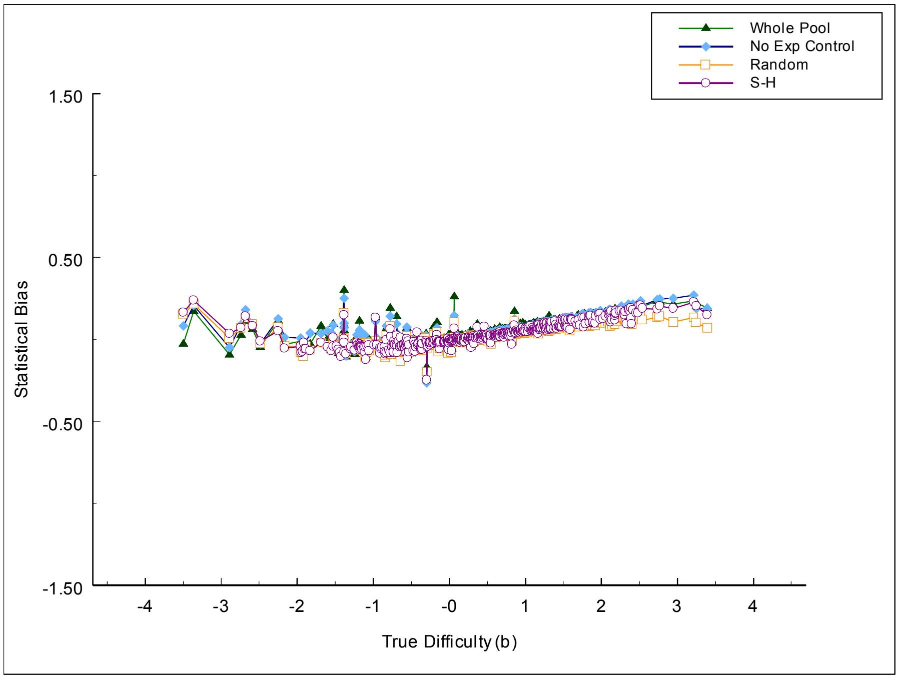

Figure 10. Bias in the *b* parameter estimates by true difficulty; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Item Pseudo-Guessing (c)

MI Estimation

MCAT Biological Sciences. The whole pool condition resulted in the largest amount of bias in the *c* parameter estimates, across all conditions. Average bias tended to be slightly positive, meaning that, on average, the estimated probability of a correct response for very low ability examinees was greater than the true probability. The highest average bias value was 0.1348, under whole pool administration with 1,000 examinees. The smallest value for average bias (0.0008) occurred under random administration with 10,000 examinees and proportional test lengths of both 1/9 and 1/12 of the pool.

Within the conditions having 1,000 examinees, differences in test length showed very little effect on bias. For the larger sample sizes, increasing test length resulted in a slight decrease in bias. Under Sympson-Hetter exposure control (10,000 examinees), average bias decreased from -0.0070 to 0.0010 when increasing proportional test length from 1/12 to 1/6.

For a given test length, changes in the number of examinees resulted in substantial effect on bias in the *c* parameter. From a sample size of 1,000 to a sample size of 10,000 bias changed from slightly positive with a wide range of values, to very close to zero (slightly negative) with much less variability. Specifically, in the case of a proportional test length of 1/12 with Sympson-Hetter exposure control, increasing sample size from 1,000 to 10,000 resulted in average bias decreasing from 0.0178 to –0.0070.

ACT Mathematics. The highest average bias in the *c* parameter (0.0237) was seen in the whole pool condition with the smallest number of examinees. Average bias values were very close to zero, ranging from slightly positive to slightly negative, depending upon the various conditions. The least amount of average bias (-0.0006) was found in the no exposure control condition with 10,000 examinees and a proportional test length of 1/9.

Very little difference in bias corresponded to changes in test length. Across the three sample sizes, increasing test length resulted in a miniscule decrease in the bias under random administration. For example, with 5,000 examinees, average bias decreased from –0.0111 to –0.0090 when proportional test length was increased from 1/12 to 1/9.

Changes in sample size caused more noticeable differences in average bias in the *c* parameter. For the longest proportional test length, increasing sample size reduced the variability of the middle half of the distribution of bias. With the shortest proportional test length, a very noticeable difference was seen in the variability of the distribution of bias estimates, and the number of outliers was greatly reduced, thus moving the overall level of bias much closer to zero. In this case, under Sympson-Hetter exposure control, increasing sample size from 1,000 to 5,000 led to a decrease in average bias from 0.0148 to –0.0004. With no exposure control and 1/9 of the pool, bias decreased from 0.0186 to -0.0006 when the number of examinees increased from 1,000 to 10,000.

EM Estimation

Figures 11 and 12 illustrate the disparity in bias between *c* parameter estimates from EM estimation and multiple imputation. For the MCAT dataset, the average EM estimates were positively biased, had larger standard deviations, and had a very wide range of values. The distinction was even more pronounced in the ACT dataset, in which the range of values in the middle 50 percent of the distribution were comparable to the range of the entire distribution (including outliers) of the bias from multiple imputation estimates. The average values of bias were very similar, and close to zero, for the two techniques, although the extreme differences in standard deviation and range make the overall results very different.

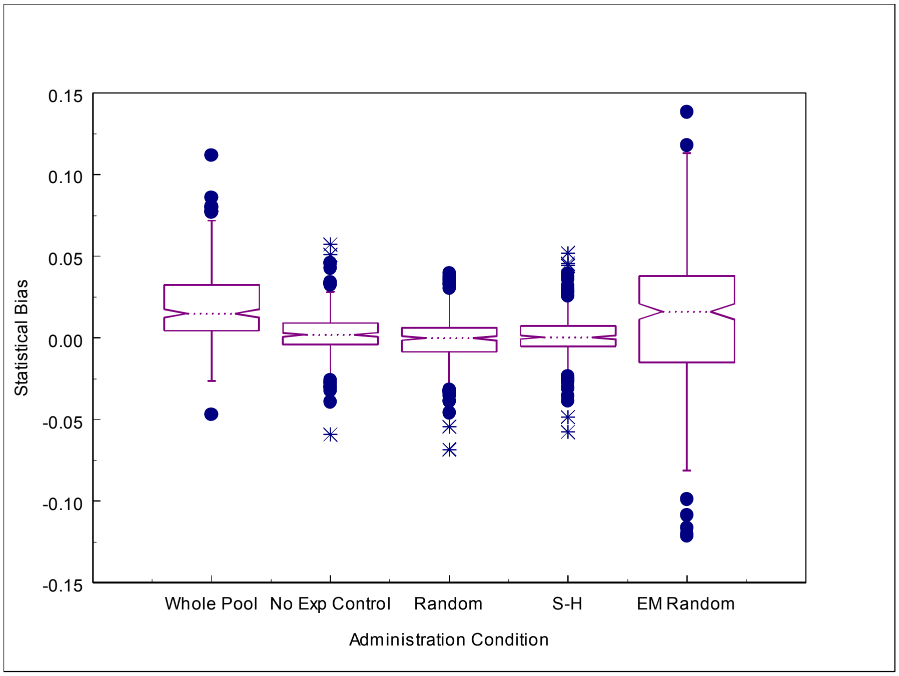

Figure 11. Bias in the *c* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

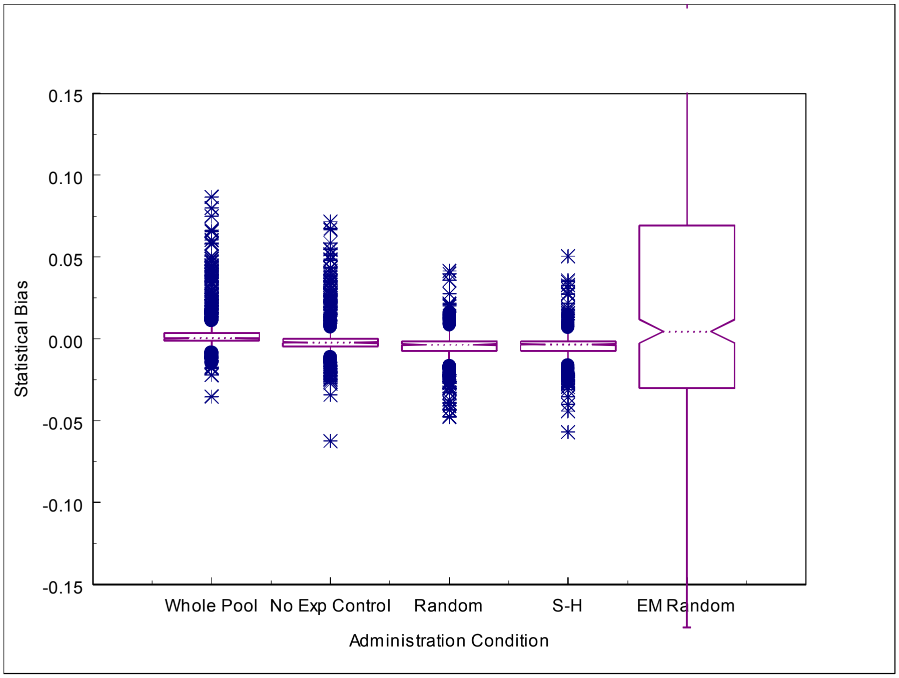

Figure 12. Bias in the *c* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Bias in *c* parameter estimates is plotted by true item pseudo-guessing in Figures 13 and 14. For all conditions in both datasets, positive bias is seen where pseudo-guessing is less, and negative bias where pseudo-guessing is greater. This results in guessing being underestimated in cases where its value is large, and overestimated when its value is small. Differences between the various item administration conditions is small, although the whole pool condition appears to have more positive bias than the others. In the ACT dataset, the no exposure control condition shows slightly higher positive bias, and the random condition slightly more negative bias. These characteristics are more pronounced in the conditions with fewer examinees and smaller proportional test lengths.

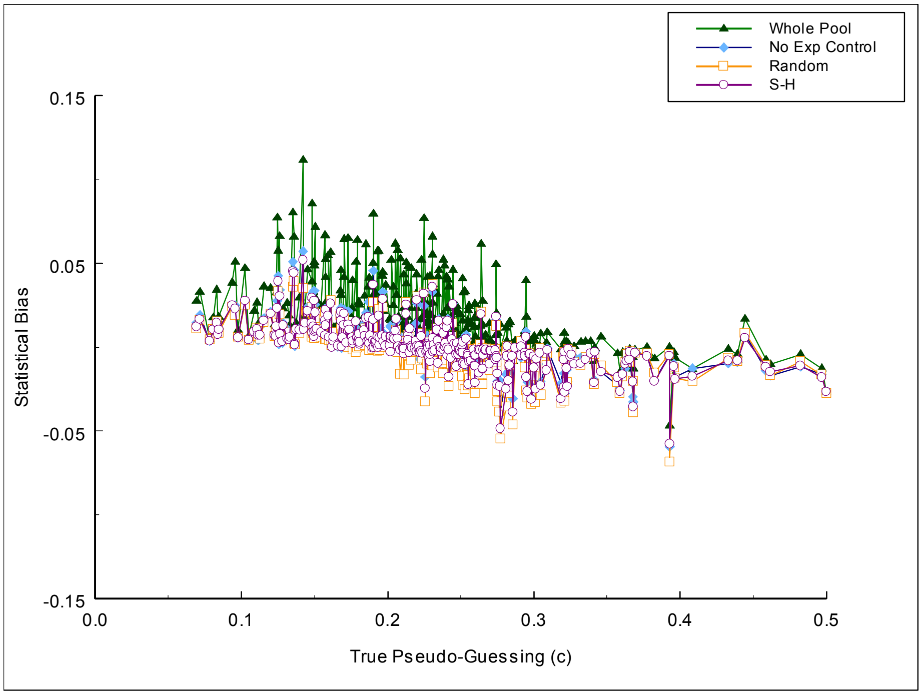

Figure 13. Bias in the *c* parameter estimates by true pseudo-guessing; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

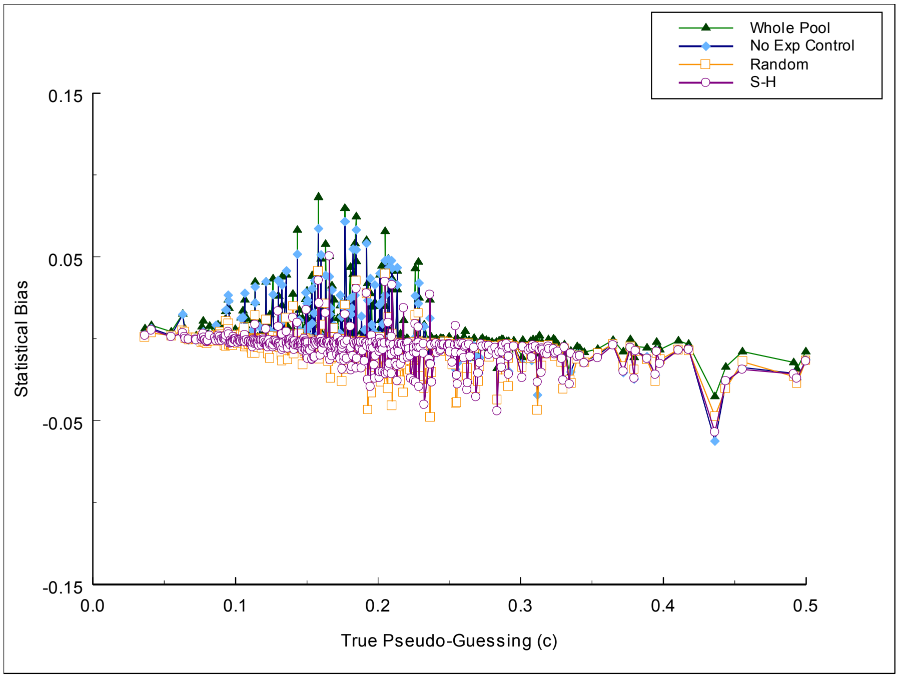

Figure 14. Bias in the *c* parameter estimates by true pseudo-guessing; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Standard Error

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.

Item Discrimination (a)

MI Estimation

MCAT Biological Sciences. The whole pool condition yielded the highest average standard error for the *a* parameter (0.0545), in the case of 1,000 examinees. All other item administration methods resulted in very similar average standard errors.

For all sample sizes, increasing the proportional test length had a negligible effect on standard error. Under Sympson-Hetter exposure control with 5,000 examinees, the average standard error was 0.0262 for 1/6 of the pool, 0.0243 for 1/9, and 0.0229 for 1/12 of the pool. These values were very similar across administration conditions.

Across all levels of proportional test length, increasing the number of examinees resulted in a corresponding reduction in the standard error of the *a* parameter estimates. This difference was most distinct when moving from a sample size of 1,000 to 5,000. For example, with a proportional test length of 1/6 under Sympson-Hetter exposure control, average standard error was 0.0529 with 1,000 examinees and 0.0262 with 5,000 examinees. Further increase in sample size to 10,000 reduced the range of outliers most noticeably under the whole pool condition. (see Figures 15 and 16).

ACT Mathematics. Larger standard errors in the *a* parameter estimate were seen in the ACT dataset. However, the range of *a* values in the ACT pool was much larger than those in the MCAT Biological Sciences pool. The whole pool with 1,000 examinees condition yielded the highest average standard error for the *a* parameter (0.1391), while the random condition had the lowest average standard error (0.0205) with 10,000 examinees and 1/12 of the item pool.

Across all sample sizes, variations in test length had very little effect on standard error in the *a* parameter estimates. For example, with no exposure control and a sample size of 10,000, standard error increased from 0.0240 with 1/12 of the pool, to 0.0253 with 1/9 of the pool, and 0.0270 with 1/6 of the pool. Similar, almost negligible, increases were seen across the other methods and conditions when proportional test length was increased.

Increase in sample size resulted in a decrease in average standard error. With a proportional test length of 1/6, and Sympson-Hetter exposure control, average standard error decreased from 0.0839 with 1,000 examinees, to 0.0377 with 5,000 examinees, and 0.0268 with 10,000 examinees. In addition to the reduction in average values, a similar reduction was found in the variability of standard errors. Under the same condition described previously, the maximum values for standard error decreased from 0.3075 to 0.2054, and finally to 0.1500, as sample size increased. Across all administration methods standard deviation, range, and number and magnitude of outliers were reduced when sample size was increased.

EM Estimation

As seen in Figures 15 and 16, standard error in the *a* parameter was much greater with EM estimation than multiple imputation. In both pools the average EM estimates contained substantially more sampling error than those obtained through multiple imputation. The average error was much larger, and the variability of the distribution of this error was distinctly greater.

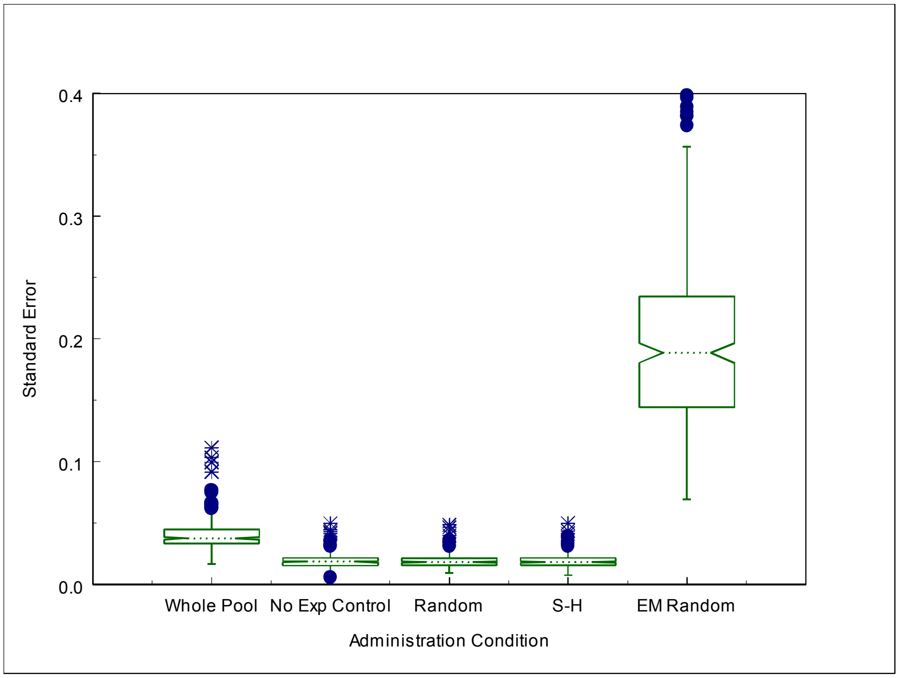

Figure 15. Standard error in the *a* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

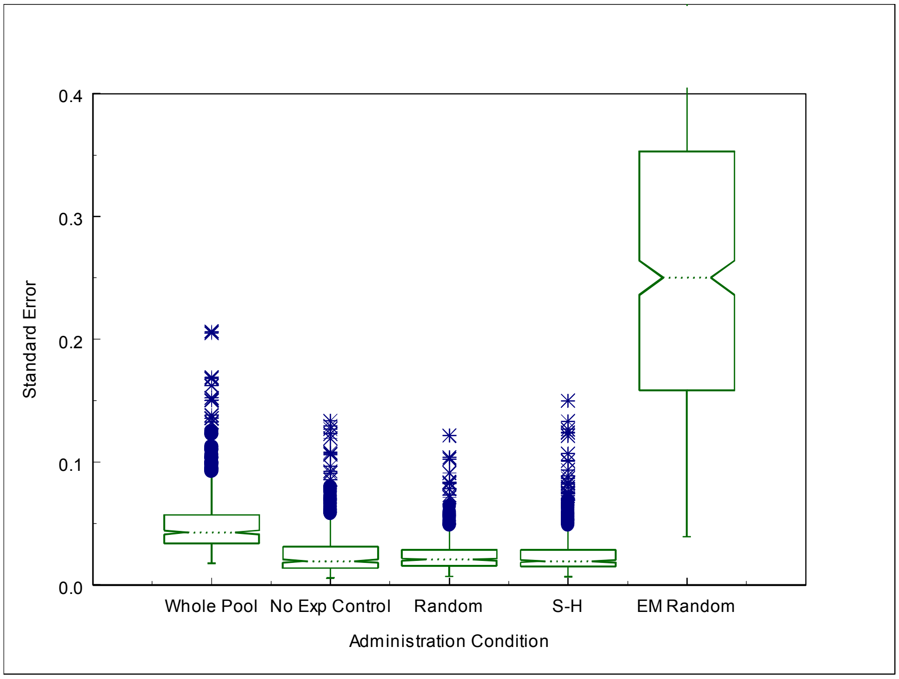

Figure 16. Standard error in the *a* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Item Difficulty (b)

MI Estimation

MCAT Biological Sciences. For the *b* parameter, whole pool administration with 1,000 resulted in the highest average value of standard error (0.1721). The smallest average standard error (0.0293) occurred under random administration with 10,000 examinees and a proportional test length of 1/12 of the pool.

For sample sizes of 1,000, variations in test length resulted in very little difference in standard error. With sample sizes of 5,000 and 10,000, an increase in proportional test length resulted in a very slight increase in average standard error. For example, with no exposure control and 5,000 examinees, average standard error was 0.0387 with 1/12 of the pool, 0.0432 with 1/9 of the pool, and 0.0516 with 1/6 of the pool.

In the case of the shortest test length, increasing the sample size led to a small but noticeable decrease in standard error on the *b* parameter estimates. For example, with random administration and a proportional test length of 1/9, average standard error decreased from 0.0690 with 1,000 examinees, to 0.0334 with 10,000 examinees. Similar patterns were found across the other two proportional test lengths and administration methods. Of note was the reduction in the magnitude of the standard error of the outliers in the whole pool condition, with the maximum values moving from near 0.65 with a sample size of 1,000 to near 0.40 with a sample size of 10,000.

ACT Mathematics. The largest average value of standard error in *b* parameter estimates (0.1393) occurred with 1,000 examinees under whole pool administration. The random condition with 10,000 examinees and a proportional test length of 1/12 resulted in the smallest average standard error (0.0235).

For smaller sample sizes, variations in test length resulted in very little change in average standard error. With a sample size of 10,000, the random condition showed virtually no change as a result of variation in proportional test length. Average standard error was 0.0235 with 1/12 of the pool, 0.0259 with 1/9, and 0.0298 with 1/6 of the pool. There was a very slight increase in the variability of standard error under no control administration as test length increased.

Increasing the sample size led to a small decrease in the magnitude and variability of standard errors on the *b* parameter estimates. The most noticeable decrease was seen when increasing sample size from 1,000 to 5,000. For example, under random administration with a proportional test length of 1/6, average standard error decreased from 0.0946 with 1,000 examinees to 0.0397 with 5,000 examinees. When sample size was further increased to 10,000, average standard error decreased to 0.0298.

EM Estimation

As seen in Figures 17 and 18, standard error in the *b* parameter was much greater with EM estimation than with multiple imputation. With both item pools, the average value of standard error in the *b* parameter estimates was substantially higher under the EM method than multiple imputation. The variability was also much greater relative to the multiple imputation results, especially with the ACT Mathematics data.

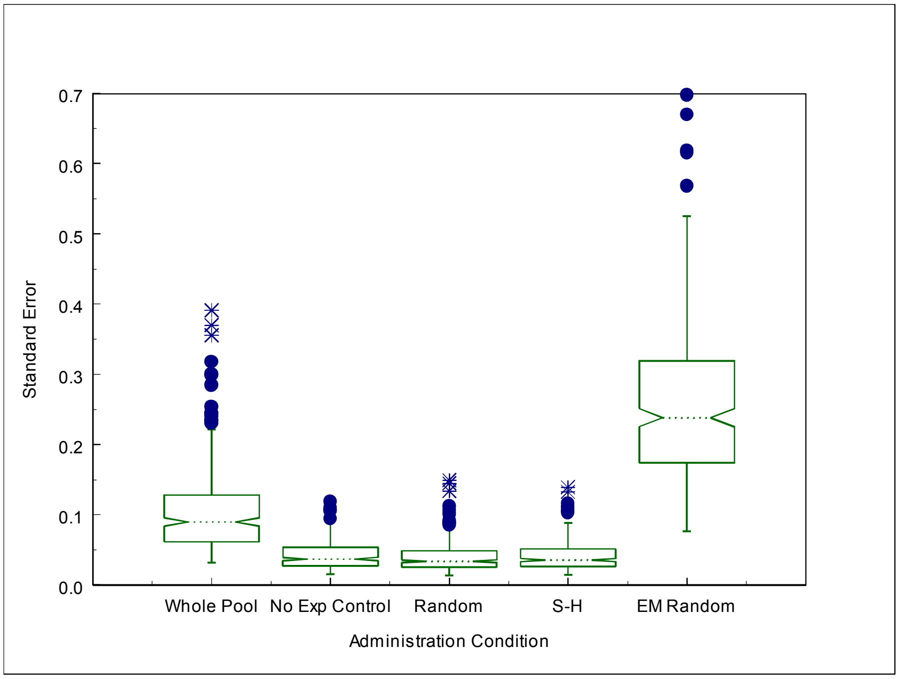

Figure 17. Standard error in the *b* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

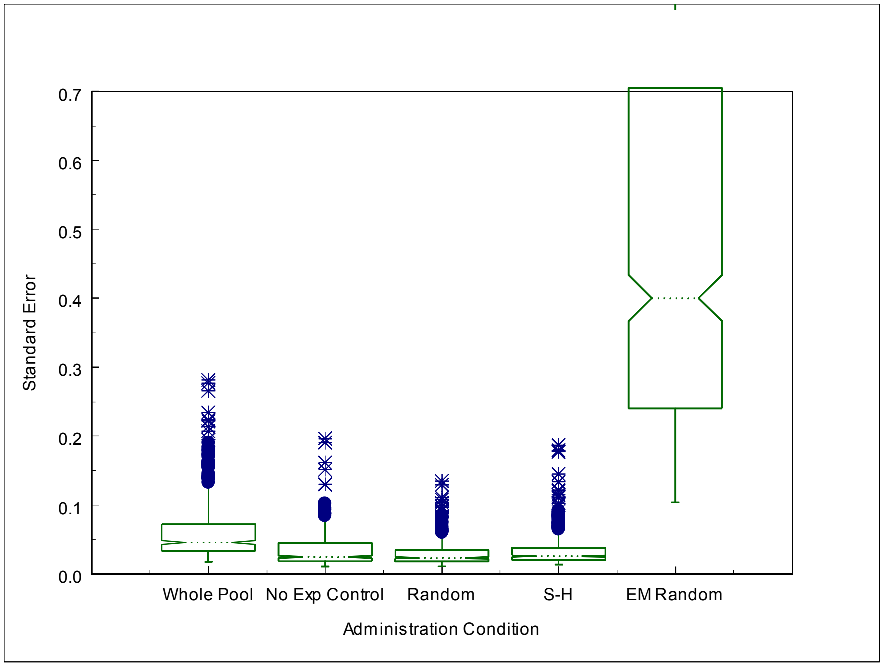

Figure 18. Standard error in the *b* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

Item Pseudo-Guessing (c)

MI Estimation

MCAT Biological Sciences. For the *c* parameter, the highest average amount of standard error (0.0420) was found under whole pool administration with 1,000 examinees. Random administration of 1/12 of the item pool with 10,000 examinees resulted in the lowest average value of standard error, 0.0101.

For all sample sizes, increasing proportional test length resulted in a very small increase in the average value and variability of standard errors. Under Sympson-Hetter exposure control with 5,000 examinees, average standard error increased from 0.0127 with the shortest proportional test length, to 0.0158 with the longest proportional test length.

Increasing the sample size led to very small reductions in average standard error in the *c* parameter estimates. Standard error was reduced most notably when moving from 1,000 to 5,000 examinees. For example, with a proportional test length of 1/6 under random administration, average standard error decreased from 0.0194 with 1,000 examinees to 0.0155 with 5,000 examinees. A smaller reduction was seen (to 0.0130) when sample size was increased to 10,000.

ACT Mathematics. For the *c* parameter, the highest average value (0.0302) of standard error occurred with whole pool administration and 1,000 examinees. The lowest average value of standard error (0.0067) was with random administration, 10,000 examinees, and a proportional test length of 1/9.

Increase in proportional test length resulted in a very small increase in average standard error. Under no exposure control administration with 10,000 examinees, average standard error increased from 0.0074 with 1/12 of the pool to 0.0091 with 1/6 of the pool. There was a slight increase in the variability of the standard errors as test length increased. For these average standard error values mentioned previously, corresponding standard deviations were 0.0066 for the shortest proportional test length and 0.0089 for the longest proportional test length.

Increase in sample size led to a small decrease in average standard error. Again, the increase from 1,000 examinees to 5,000 examinees resulted in the more noticeable standard error decrease. In the case of Sympson-Hetter exposure control with a proportional test length of 1/9, average standard error decreased from 0.0131 with 1,000 examines, to 0.0088 with 5,000 examinees, and 0.0076 with 10,000 examinees.

EM Estimation

As seen in Figures 19 and 20, item calibration using the EM algorithm led to much larger standard errors in *c* parameter estimates than with multiple imputation. Both the magnitude of the average values and the range of variability of these values was substantially higher when using the EM method. This held true across both item pools.

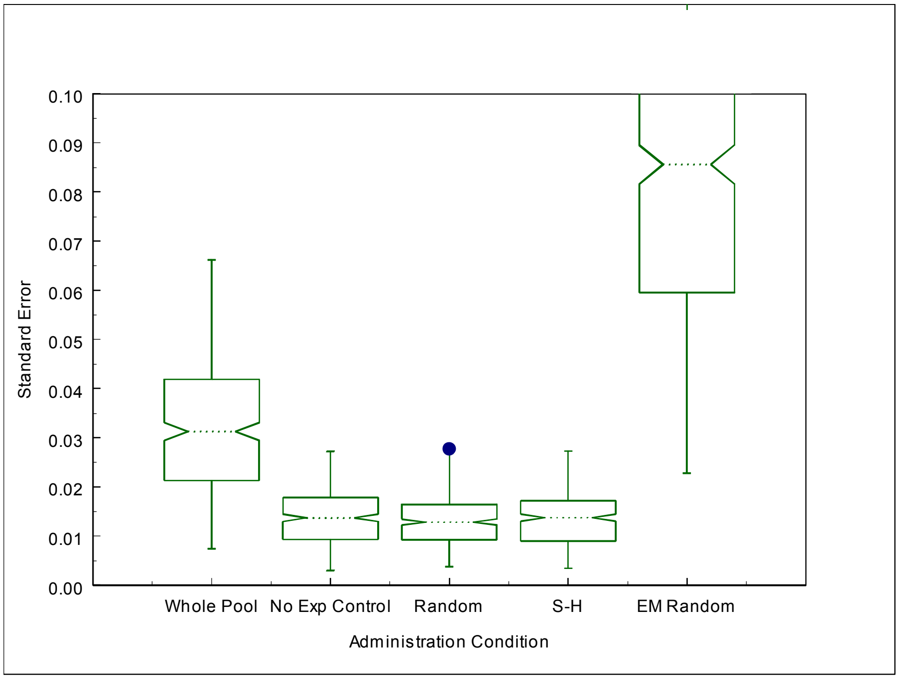

Figure 19. Standard error in the *c* parameter estimates; MCAT Biological Sciences, 1/6 of 312 items; 10,000 examinees.

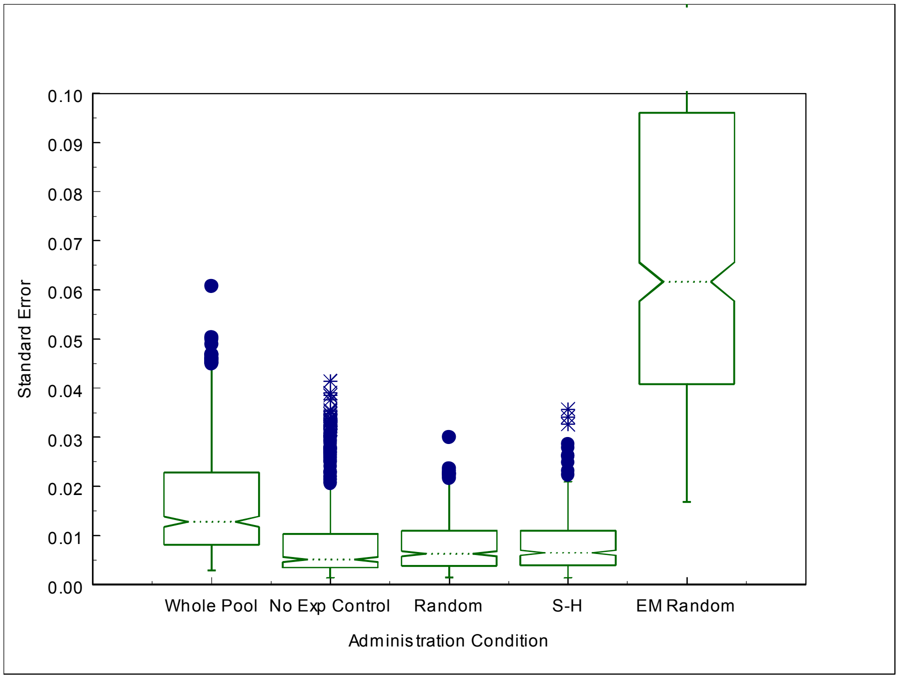

Figure 20. Standard error in the *c* parameter estimates; ACT Mathematics, 1/6 of 480 items; 10,000 examinees.

DISCUSSION

The conditions examined in this study represented average values from 83% to 92% missing data in the item response matrices. In none of the cases was BILOG able to successfully calibrate the sparse response data matrices without application of missing data treatments. This emphasizes the need for an alternate solution. 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 multiple imputation technique resulted in the lower amount of statistical bias in item parameter estimates, as compared to parameter estimates from ML estimation with the EM algorithm. Variations in proportional test length had a small effect on bias. In general, longer proportional test lengths resulted in smaller amounts of bias across item parameter estimates. Variations in sample size had a much more pronounced effect on the levels of bias. As the number of examinees increased, the amount of bias generally decreased, and the variability across the distribution of bias and magnitude of outliers decreased.

The multiple imputation technique also resulted in the lower amount of standard error in item parameter estimates, as compared to parameter estimates from ML estimation with the EM algorithm. Variations in proportional test length led to very small changes in average standard error. There was a tendency for average values to decrease as proportional test length decreased. As sample size increased standard errors in the item parameter estimates showed a small decrease. The whole pool condition resulted in generally higher values of standard error in item parameter estimates. Standard errors of item parameter estimates obtained via calibration with the EM method were substantially higher and had a much wider range of variability than did those obtained through multiple imputation.

The relatively small levels of statistical bias and sampling error underscore the potential of multiple imputation in applied testing programs. In this study MI performed well across two datasets that differed greatly, both in terms of item pool distributions and item administration constraints. The one area of concern arose with the smallest sample size (1,000). There were a few cases with the MCAT Biological Sciences pool, and many more cases with the ACT Mathematics pool, in which datasets that had been filled in with imputations could not be calibrated by BILOG. In those cases, however, once new imputations were performed, the problem was remedied.

The ML estimation with the EM algorithm technique does not appear to be as promising for response data matrices with a small number of examinees and a large amount of sparseness. It is possible that the conditions in this study were too demanding, i.e., too much sparseness and too many items that only had data on a few examinees. In Ban et al., (2001) application of the EM algorithm to the calibration of pretest items included several differences that may have contributed to their more positive results. Because the focus of that study was pretest items, the researchers were dealing with a much smaller number of items that needed to be calibrated (as opposed to an entire item pool). These pretest items were combined with a set of previously calibrated items and the entire set was then sent through the calibration process together. In this situation, the parameter values of the non-pretest items were held constant. This should have contributed substantially to the stability of the parameter estimates of the pretest items—pulling them back to reasonable values.

Wainer and Mislevy (1990) suggested the use of the EM algorithm with one iteration (sometimes referred to as OEM) as a method for calibrating sparse data. Ban et al. (2001) tested this approach, as well as others, and found that using multiple EM iterations produced parameter estimates with lower amounts of bias, standard error, and RMSE. In effect, the OEM would be the same as conducting a single imputation and then recalibrating. The use of multiple imputation is a far stronger choice, especially given the potential for sampling error.

The performance of the EM algorithm was markedly different across the two datasets used in this study. The best performance was seen with the MCAT Biological Sciences dataset. This item pool was characterized by a very constrained distribution of *a* parameter values (from 0.25 to 1), and a fairly even distribution of *b* parameters between –2 and 2. In contrast, the ACT pool contained items with *a* parameter values ranging from 0.25 to 2.5. In an effort to realistically simulate the MCAT Biological Sciences test administration, items were administered according to a passage-based algorithm. Items in the ACT Mathematics pool were discrete, i.e., not subject to passage-based administration constraints. The restricted range of variability on the *a* parameter may have helped keep the variability of EM estimates lower with the MCAT data. These two factors (item pool distribution and administration constraints) were likely contributors to the differences in the performance of the EM algorithm across the two datasets.

LIMITATIONS

The results of this study should be interpreted within the context of the study's limitations. Simulation studies provide a real opportunity to investigate psychometric models within a CAT environment. Only when truth is known (i.e., true examinee ability and true item parameters) can experimental conditions such as those in this study be accurately executed and evaluated. However, the use of simulation to generate data also restricts the degree to which inferences can be made to the operational testing environment. Although real examinee responses were used in the initial stage of the project to provide a baseline for comparison, simulation of examinees in the computerized testing environment cannot completely substitute for actual examinees.

In an operational testing environment, it would not be possible to use the "true" *a*, *b*, and *c* parameter values for generating multiple imputations. In such a setting, the imputations would be based on item parameter estimates, and thus the new item parameter estimates that would be calibrated from this imputed data would contain more error than those based on true item parameters.

Finally, the whole pool condition was included as a reference distribution. It simulates the full data condition that would occur in a case such as a computerized fixed test in which all examinees would be presented with a full set of items. However, it would not be a realistic option in an operational setting, as it would never be possible to give all examinees all the items in the pool. Thus, care should be taken in interpreting the performance of the whole pool condition as an option for operational CBT.

SUGGESTIONS FOR FURTHER STUDY

Results of this study provide information regarding the problem of sparseness for CAT item calibration, and give a starting point for further analysis and modeling of sparseness and its effects on IRT item calibration. The technique of ML with EM has been previously applied to IRT item calibration, but had not been tested under conditions with such a high degree of sparseness as occurred in this study. This approach requires further research and simulation in order for it to become a useful and practical approach in rigorous applications such as this. It appears to be more promising for situations in which fewer items have large amounts of sparseness, and data are available on large numbers of examinees. The multiple imputation technique has not previously been applied to the problem of sparseness in IRT item calibration. This initial trial has shown it to have potential. The results from this study have identified an additional tool for psychometricians to use in dealing with the complex problem of online item recalibration. The simplicity and relative robustness of multiple imputation make it a promising avenue for further study.

In addition to the study of missing data treatments as a remedy for sparseness resulting from adaptive tests, it may be worthwhile to investigate alternative methods for CAT administration for calibration purposes. Operational strategies might include seeding items to appear in the same position across many examinees or administering sets of items to groups of examinees. Seeding items would help to ensure data from large numbers of examinees representing a wide range of ability. Administering sets of items would be similar to passage-based or testlet administration (see Wainer & Kiely, 1987), and would result in more commonly co-occurring items. Multi-stage testing is a strategy similar to testlets, and may help mitigate calibration problems (see Reese, Schnipke, & Luebke, 1997; Schnipke & Reese, 1999).

EDUCATIONAL IMPORTANCE OF THE STUDY

In the process of investigating issues related to conversion of the existing paper-and-pencil testing programs to CAT administration, examining the severity and the consequences of sparseness and determining the extent of errors in calibration such sparseness might cause are important issues. 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.

REFERENCES

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.

Davey, T., Nering, M.L., & Thompson, T. (1997, March). *Realistic simulation of item response data.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Harwell, M., Stone, C.A., Hsu, T.C., & Kirisci, L. (1996). Monte Carlo studies in item response theory. *Applied Psychological Measurement 20*, 101-125.

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.

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.

Mislevy, R.J. & Bock, R.D. (1990). BILOG 3: Item analysis and test scoring with binary logistic models. [Computer software and manual]. Chicago: Scientific Software.

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.

Pommerich, M. (2002, April). *The effect of administration mode on test performance and score precision, and some factors contributing to mode differences*. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.

Reese, L., Schnipke, D., & Luebke, S. (1997, March). *Incorporating content constraints into multi-stage adaptive testlet design.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

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.

Rubin, D.B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Schafer, J.L. (2001). Multiple imputation FAQ page. Retreived April 2001, from http://www.stat.psu.edu/~jls/mifaq.html

Schnipke, D.L. & Reese, L.M. (1999). A comparison of testlet-based test designs for computerized adaptive testing. (LSAC Computerized Testing Report 97-01). Newtown, PA: LSAC.

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.

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

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.