Center for Advanced Studies in
Measurement and Assessment

CASMA Research Report

Number 39

Evaluation of Comparability of Scores and Passing Decisions for Different Item Pools of Computerized Adaptive Examinations

Wei Wang, Shichao Wang, Michael J. Kolen,
Won-Chan Lee, Jaime L. Peterson,
Mengyao Zhang, and Seohong Pak

The University of Iowa

November 2014

Center for Advanced Studies in
Measurement and Assessment (CASMA)
College of Education
University of Iowa
Iowa City, IA 52242
Tel: 319-335-5439
Web: www.education.uiowa.edu/casma

Contents

List of Tables

List of Figures

Abstract

Many computerized adaptive testing (CAT) programs employ multiple item pools and rotate them in use in order to address security concerns. To ensure the fairness of the assessment, scores and passing decisions across different item pools should be psychometrically comparable. Only guidelines have been provided in the previous research on how to evaluate the comparability of different item pools of a CAT examination. This study assessed psychometric comparability of multiple item pools for two operational CAT examinations. In addition, detailed computational procedures were provided for the statistics that are used for assessing such comparability. The results showed that for each examination, the scores and passing decisions were psychometrically comparable across the four different pools.

1 Introduction

Computer adaptive testing (CAT) based on item response theory (IRT) is being used to deliver tests in many situations as an alternative to paper-pencil testing. Among examples of CAT examinations are the Graduate Management Admission Test `http://www.mba.com/us`, the National Council Licensure Examinations for Registered Nurses and for Practical and Vocational Nurses `https://www.ncsbn.org`, and the CAT version of the Armed Services Vocational Aptitude Battery `http://www.military.com/join-armed-forces/asvab`. CAT has become popular and more testing programs have become interested in it, because it offers various measurement advantages over paper-pencil (P&P) testing, such as improved examinees' testing experience, increased testing efficiency, convenient test scheduling, and immediate scoring and feedback to examinees (Wainer, 2000).

In CAT, often, due to a large volume of test takers, over-use of an item pool might occur when continuously administering tests from the same pool, which could further cause test security concerns. To address test security, in addition to controlling item exposure rate of a single pool, many CAT programs also develop multiple item pools at the same time or have several item pools available, and then rotate them in use (Davey & Nering, 1998; Wang & Kolen, 2001). In this approach, multiple item pools of the same test are developed, and each contains different items. Each pool is used for a certain period of time or for a certain number of testing administrations. Once the limits are reached, the old item pool is replaced by a new one. Through rotating multiple different item pools, the purpose of reducing the over-use of item pools can be achieved.

When examinees are administered CAT examinations which use multiple item pools, they expect their scores to be comparable regardless of the item pool from which their test is drawn. However, each of the pools contains different items. In addition, although all pools are built to the same technical specifications, there will be some differences in the statistical characteristics of the items in the different pools. For these reasons, it is of high importance to evaluate the extent to which the pass-fail decisions and proficiency estimates are comparable, psychometrically, across pools. Wang and Kolen (2001) provided criteria that can be used to assess the comparability of scores across pools.

Wang and Kolen (2001) referred to psychometric property/reliability criteria. These criteria include conditional equity criteria and overall criteria. Equity criteria are based on Lord's (1980) concept of equity. In the context of computerized adaptive testing with multiple item pools, the equity criteria imply that examinees should be indifferent to the pool from which their test is drawn. From a psychometric perspective, the equity criteria implies that each examinee should be expected to earn the same score regardless of the pool used, which is referred to as first-order equity. The equity criteria also imply that an examinee should be measured with the same precision regardless of which pool is used, which is referred to as second-order equity. Besides the first-order equity and second-order equity, passing-score equity should also be considered. Passing-score equity is referred to as "equal probabilities of achieving passing scores"

(Wang & Kolen, 2001). Passing score equity implies that for examinees of a given proficiency, the probability of passing is the same regardless of the pool used to assess the examinee.

Wang and Kolen (2001) also referred to overall criteria, which are for the whole population of examinees (rather than conditional on examinee proficiency as with the equity criteria). The same distributions criterion implies that for a given population of examinees, the score distribution is the same regardless of the item pool used. For tests with passing scores, this criterion implies that the same proportion of examinees in a given population would pass, regardless of the pool used. The same reliability criterion, another overall criterion, implies that, for the population of examinees, the reliability of scores (i.e., ability estimates) is the same regardless of the pool used. For tests that have passing scores, this criterion also implies that for the population of examinees the scores on each pool will have the same decision consistency coefficient regardless of the pool used. Decision accuracy indices including false positive and false negative error rates will also be the same regardless of the pool used.

In addition, as Wang and Kolen (2001) described, it is unnecessary to wait to perform this comparability study after these pools are used operationally. Instead, the study can be conducted at an early stage of CAT development by using simulation technique.

As described earlier, when multiple item pools are in use for a CAT examination, there is a concern that scores earned on different pools might not be strictly comparable. Comparability can be strengthened by developing the pools strictly following content specifications and statistical specifications. However, even when all pools are built to the same specifications, the pools might still differ in statistical characteristics, which could cause incomparability of scores. To ensure the fairness of the assessment, it is important and necessary to evaluate comparability of item pools psychometrically. In addition, such evaluation is an important piece of validity evidence for any CAT examination which uses multiple item pools. Most of the research related to comparability issues in CAT focuses on assessing the comparability of CATs and P&P tests (e.g., Davey & Thomas, 1996; Stocking, 1994; Thomasson 1997; van der Linden 2006; Wang & Kolen, 2001). Little attention has been paid to the psychometric comparability of multiple item pools used in a CAT testing program, and very little research has addressed this issue (Wang & Kolen, 2001). No empirical studies have been conducted to illustrate how to assess psychometric comparability of different CAT pools. In addition, no computational procedures have been provided regarding how to calculate the statistics that were suggested in Wang and Kolen (2001).

The primary purpose of the present study is to illustrate how to assess psychometric comparability of scores from operational CAT pools. In addition, in the context of a simulation study, the second purpose of the present study is to provide detailed computation of various evaluation criteria that are used to assess comparability of different pools. In this study we assessed the psychometric comparability across four item pools for each of two licensure testing programs. Comparability was assessed using simulation techniques. To conduct the simulation, we received from the two testing programs item parameter estimates for items in each of these four pools, indicators of the item content category for the items, and distributions of examinee ability estimates. In addition, we received information from the testing programs on how the adaptive tests are administered (e.g., content balancing, stopping rules, proficiency estimation procedures, passing scores) so that our simulation closely mirrored the psychometric procedures used operationally.

According to the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014), Standard 5.12, "a clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably." The study performed in this paper is intended to help provide the evidence suggested by the Test Standards for any CAT examination that uses multiple item pools in a rotating manner or at the same time.

2 Methodology

To assess the comparability of multiple item pools, simulation studies were conducted using information provided by the testing programs. In this section, the data sources for the simulation study are provided and the CAT simulation system is described. Then, simulation parameters are presented in detail. Finally, evaluation criteria are addressed.

2.1 Data Source and Description of the CAT Simulation System

In this study, comparability of multiple item pools was assessed for two variable length computerized adaptive tests, both scaled using the Rasch model (Rasch, 1960). The two CAT examinations are from two different operational licensure testing programs. For convenience, one examination is referred to as Test A, and the other one is referred to as Test B. For both examinations, the minimum test length is 60. The maximum test length allowed is 250 items for Test A and 180 items for Test B. The examinations each have a single passing standard. The cut score (on theta metric) is -0.11 for Test A and -0.32 for Test B. Pass-Fail status is reported to examinees.

For each examination, four item pools are available, and they contain different items from each other. These pools are rotated quarterly. The item parameter estimates for items in each of these four pools, indicators of the item content category for the items, and distributions of examinee ability estimates were provided by each testing program.

The two testing programs use the same stopping rules and theta estimation methods. For both tests, at the start of the administration, Owen's Bayesian method (Owen, 1969) is used for ability estimation until at least one item correct and one item wrong are obtained. Then, the method for ability estimation is switched to the maximum likelihood estimation (MLE) approach. Prior to the maximum number of items permitted by the system being reached, for both tests, one-tail 95% confidence interval rule is used to determine whether the candidate should continue the test. If a candidate's ability estimate ($\widehat{\theta}$) is 1.65 times the standard error (of $\widehat{\theta}$) greater than the cut score, the system stops delivering items to the examinee, and he/she passes the examination. If an ability estimate for a candidate plus 1.65 times the standard error is less than the cut score, no more items are administered to the candidate, and he/she fails the test. Otherwise, the examination continues until the maximum number of items is reached. If a clear decision cannot be made by the time a candidate has completed the maximum number of items, the pass/fail decision is made based on the ability estimate after the last item. If a candidate's final ability estimate is equal to or greater than the cut score, he/she passes the examination; otherwise, he/she fails the examination.

A simulation system, written in programming language C, was developed to mirror the operational CAT systems used to deliver the two examinations. The CAT simulation system was consistent with the operational CAT systems in terms of content balancing, ability estimation methods, classification rules, and stopping rules.

2.2 Simulation Parameters

Eighty-one equally spaced theta (ability) points from -3 to 3 with an increment of 0.075 were used. For each item pool and for each examination (four pools for Test A and four pools for Test B), simulation studies were conducted using the 81 theta points and the item parameters provided by each of the two testing programs. The same procedures were replicated 1,000 times for each theta point using each pool of the two examinations.

For each pool of the same examination, the simulation was performed using two different seed numbers used to generate random numbers. Results obtained from two different runs of the same pool were compared. The purpose of this analysis was to show whether the differences among pools are larger than the differences when a single pool is run twice with different seed numbers for generating random numbers.

2.3 Evaluation Criteria

One important criterion for assessing psychometric comparability of multiple item pools is whether, across different pools, items in the same content category have the same characteristics. Therefore, once item parameters were obtained from the testing programs, for the same examination (Test A or B), summary statistics for item difficulty parameter estimates (bs) needed to be reviewed including the number of items, minimum value, maximum value, mean value, and standard deviation. These statistics were computed for each content category within a pool and for the whole pool. In addition, based on item response theory, information functions were computed for each content category within a pool and for each full pool.

After the simulation studies were completed, the range and mean percentages of the numbers of items from each content category of each pool were calculated over 81,000 simulated examinees (81 theta values × 1,000 replications). In addition, several conditional and marginal statistics were computed based on the simulation results in order to evaluate the psychometric comparability of item pools for each examination. The computation of these conditional and marginal statistics is introduced in detail in the following section.

2.3.1 Conditional Statistics

The conditional statistics considered in this study include conditional bias (CBIAS), conditional standard error (CSE), conditional mean squared error (CMSE), average number of items taken at each theta point, and conditional passing rate. Equations for these conditional statistics are provided below.

Conditional Bias (CBIAS) Conditional bias (CBIAS) is defined as the difference between the mean of the estimated theta values and the true theta. The equation to compute CBIAS is as follows:

$$CBIAS_j = \overline{\widehat{\theta}}_j - \theta_j, \tag{1}$$

and

$$\overline{\widehat{\theta}}_j = \frac{\sum_{r=1}^{1000} \widehat{\theta}_{jr}}{1000}, j = 1, 2, ..., 81, \tag{2}$$

where $\overline{\widehat{\theta}}_j$ is the mean ability estimate over 1,000 simulated examinees for a particular ability; j is an index for theta values; r is an index for replications; θ_j is the true theta value; and $\widehat{\theta}_{jr}$ is the ability estimate of θ_j on the rth replication.

Conditional Standard Error (CSE) of Estimated Theta There are two approaches to computing conditional standard error (CSE) of an estimated theta. In the first approach, CSE is computed as the standard deviation of the 1,000 estimated thetas for a particular true theta value, and hereafter, this CSE is called empirical conditional standard error (ECSE). The equation used to compute ECSE can be expressed as:

$$ECSE_j = \sqrt{\frac{\sum_{r=1}^{1000} \left(\widehat{\theta}_{jr} - \overline{\widehat{\theta}}_j\right)^2}{1000 - 1}}. \tag{3}$$

In the second approach, the computation of CSE is based on the test information function for the set of items administered to the simulated examinee, and this CSE is named as analytical conditional standard error (ACSE). ACSE is computed using

$$ACSE_j = \frac{1}{1000} \sum_{r=1}^{1000} SE\left(\widehat{\theta}_{jr} | \theta_j\right), j = 1, 2, ..., 81, \tag{4}$$

where

$$SE\left(\widehat{\theta}_{jr}|\theta_j\right) = \sqrt{\frac{1}{\sum_{i=1}^{n} P_i(1-P_i)}}, j = 1, 2, ..., 81. \tag{5}$$

In Equation 5, n is an index for a number of items administered, and P_i is the probability of an examinee with ability of θ_j correctly answering item i. To compute the standard error for each replication in this simulation, estimated theta ($\widehat{\theta}_{jr}$) on replication r was used in Equation 4 instead of true theta. Then, conditional on each true theta value, the average standard error is computed over the 1,000 replications, which is the ACSE.

Conditional Mean Squared Error (CMSE) The conditional mean squared error (CMSE) is the sum of squared CBIAS and squared CSE. Since the CSE can be computed in two different ways, two different CMSEs are computed. The CMSE based on the ECSE is referred to as the empirical conditional mean squared error (ECMSE), and the CMSE based on the ACSE is referred to as the analytical conditional mean squared error (ACMSE), which are given by

$$ECMSE_j = CBIAS_j^2 + ECSE_j^2, j = 1, 2, ..., 81, \tag{6}$$

and

$$ACMSE_j = CBIAS_j^2 + ACSE_j^2, j = 1, 2, ..., 81. \tag{7}$$

Average Number of Items Taken at Each Theta Point Conditional on each true theta, the average number of items administered over 1,000 replications was computed. The results are presented using plots in a later section.

Conditional Passing Rate Conditional on each true theta, the percent of times out of 1,000 replications that the CAT simulation system produced a pass decision was computed as the conditional passing rate. Graphs are used to present the results.

2.3.2 Marginal Statistics

In addition to the conditional statistics, several marginal statistics were also computed for each pool, including reliability, decision accuracy, decision consistency, and numbers of administered items for theta intervals (defined by each testing program), and standard errors of estimated theta for theta intervals. When computing the marginal statistics, except for decision consistency across pools as discussed later, two different types of weights were used for each pool. The first type is weights computed from an empirical distribution which were obtained based on the actual numbers of examinees taking each item pool and their theta estimates in operational administrations. These weights are referred to here as empirical weights. The second type is weights computed from a normal distribution, which has the same mean and standard deviation as the empirical distribution, and these weights are referred to here as normal weights.

The 81 true thetas used in the simulation have different empirical weights and normal weights across pools and examinations.

The layout for true thetas, estimated thetas, and weights are presented in Table 1. Each true theta (the second column from the left) has a weight (the leftmost column) associated with it, which can be either empirical or normal. Because the simulation was performed 1,000 times, each true theta has 1,000 associated estimates (the third column), and each theta estimate's weight accounts for 0.1% of the corresponding true theta's weight. The notation in Table 1 is used in equations for computing the marginal statistics as described below.

Reliability Reliability is calculated on the theta scale. The steps to compute reliability for each pool are:

1. Compute the variance of the theta estimates ($\sigma^2_{\widehat{\theta}}$).

$$\sigma^2_{\widehat{\theta}} = \frac{1}{1000-1} \sum_{j=1}^{81} \sum_{r=1}^{1000} w_j \left(\widehat{\theta}_{jr} - \overline{\widehat{\theta}} \right)^2 , \tag{8}$$

   where

$$\overline{\widehat{\theta}} = \frac{1}{1000} \sum_{j=1}^{81} \sum_{r=1}^{1000} w_j \widehat{\theta}_{jr}. \tag{9}$$

   In Equations 8 and 9, w denotes the weight for each true theta.

2. Compute the error variance (σ^2_E). At first, conditional on each true theta, compute the variance of its 1,000 theta estimates (Equation 10).Then, compute the weighted average of the conditional variances across the 81 true thetas, which is the error variance (Equation 11).

$$\sigma^2_{\widehat{\theta}_j} = \frac{\sum_{r=1}^{1000} \left(\widehat{\theta}_{jr} - \overline{\widehat{\theta}}_j \right)^2}{1000-1}. \tag{10}$$

$$\sigma^2_E = \sum_{j=1}^{81} w_j \sigma^2_{\widehat{\theta}_j}. \tag{11}$$

3. Reliability is defined as:

$$Reliability = 1 - \frac{\sigma^2_E}{\sigma^2_{\widehat{\theta}}}. \tag{12}$$

Decision Accuracy (DA) When the true theta is available, the fail/pass decision can be made by comparing the true theta to the cut score. An accurate decision is obtained when the decision made based on the estimated theta in the simulation is consistent with the decision made based on the true theta. For each pool, decision accuracy (DA) is defined as the weighted average proportion of times that the estimated ability leads to the same decision as the true theta. The equation for DA is:

$$DA = \sum_{j=1}^{81} w_j p_j, \tag{13}$$

where p_j is the percent of times out of 1,000 replications that the same pass/fail decisions are made based on theta estimates and the true theta values.

Decision Consistency (DC) Decision consistency (DC) was computed in two different ways: (1) based on a single pool (SDC) of an examination, and (2) across different pools (MDC) of the same examination.

To compute SDC, the steps are as follows:

1. Conditional on a true theta value (θ_j), compute the proportion of times out of 1,000 replications that a pass decision is made (q_j). Then, the proportion of times that a fail decision is made is computed as $1 - q_j$;
2. Suppose that an examinee with a true theta value, θ_j, takes the same single pool twice. The probability that the examinee passes the examination twice is calculated as $q_j \times q_j$, and the probability of failing the examination twice is $(1 - q_j)^2$;
3. Conditional on θ_j, sum the two probabilities calculated in Step 2, and the sum is referred to as z_j; and
4. Use Equation 14 to compute SDC:

$$SDC = \sum_{j=1}^{81} w_j z_j, \tag{14}$$

where w_j is the weight of a true theta θ_j, and is defined as before.

MDC was calculated in a way similar to SDC. Suppose examinees take pool 1 and pool 2 of the same examination. The steps to compute MDC are:

1. Conditional on a true theta value (θ_j), for pool 1, compute the proportion of times out of 1,000 replications that a pass decision is made (q_{1j}), and then, the proportion of times that a fail decision is made is computed as $1 - q_{1j}$. Do the same calculation using pool 2, and the proportion of times out of 1,000 replications that a pass decision is q_{2j} and the proportion of times that a fail decision is made is $1 - q_{2j}$;
2. Suppose that an examinee with a true theta value, θ_j, takes examinations that are constructed from pool 1 and pool 2 respectively. The probability that the examinee passes the two pools is calculated as $q_{1j} \times q_{2j}$, and the probability to fail the two pools is $(1 - q_{1j}) \times (1 - q_{2j})$;
3. Conditional on θ_j, sum the two probabilities calculated in Step 2, and the sum is referred to as z_j; and

4. Use Equation 14 to compute MDC. It is important to note that the term w is defined in a different way for MDC. Two types of weights, w, were used for computing MDC. One set of weights were obtained based on an empirical distribution which was the average of the four Test A or Test B empirical distributions. The other set of weights were obtained based on a normal distribution with the average mean and the average standard deviation of the four Test A or Test B normal distributions.

Average Number of Items Taken and Standard Errors of Theta by Theta Intervals For convenience of reporting, testing programs often group individual estimated theta values into several intervals and report results for those intervals. An example of theta intervals is shown in Table 2, which is provided by the two testing programs. These intervals are grouped based on theta estimates instead of true thetas. Average number of items taken and standard errors of estimated theta are often reported for theta intervals.

The steps to compute the average number of items taken by theta intervals were as follows: First, theta estimates were grouped into theta intervals; second, within each theta interval, the mean number of items administered was calculated at each true theta value; third, the weighted mean number of items within each interval was obtained, conditional on true theta; and fourth, for each interval, the sum of the weighted mean number of items was divided by the sum of the weights to obtain the marginal average number of items taken.

The steps to compute standard errors of theta by theta intervals were similar to those for computing the average number of items taken by theta intervals as follows: First, theta estimates were grouped into 11 theta intervals; second, within each theta interval, the mean ACSE was calculated at each true theta value; third, the weighted ACSE within each interval was obtained, conditional on true theta; and fourth, for each interval, the sum of the weighted ACSE was divided by the sum of the weights to obtain the marginal standard error.

3 Results

In this section, results are reported for both examinations (Test A and Test B). In general, the four item pools of each examination are psychometrically comparable based on the evaluation criteria described in the Methodology section. In addition, the main findings are highly consistent across different pools and different testing programs.

3.1 Summary Statistics of Item Difficulties and Information for Contents and Pools

Tables 3 and 4 provide the summary statistics of item difficulties in each pool and in each content category within a pool for Test A and Test B, respectively. Both tables provide results of the number of items, minimum b value, maximum b value, mean value of b, and standard deviation of bs. According to Table 3, the four pools for Test A appear to be constructed similarly that the numbers of items in each content category are very close across pools, and also the magnitudes of the other four statistics are similar across pools. The same finding is observed for the pools for Test B (see Table 4).

Figure 1 illustrates the comparison of pool information functions, which are the information functions for the entire pool, among the four Test A item pools. The solid line denotes Pool 1, the long dashed line is for Pool 2, the median dashed line for Pool 3, and the dotted line for Pool 4. The representations of these lines are also used for Test B, and they are consistent throughout this paper (except Figures 25 and 26). The overlapping of the four curves shows that the four item pools produce very similar information along the theta scale, which indicates the comparability of the item pools of Test A. In addition, the information for each content category in Test A was compared across the four pools, and the results are shown in Figure 2, which contains eight panels, one for each content category. In each panel, the four curves almost overlap with each other. The comparisons demonstrate that the four item pools have similar information considering each content category of Test A.

Figure 3 shows the comparison of pool information among the four Test B item pools, and Figure 4 shows the comparison of information among the pools when each content category in Test B is considered. The overlapping of the four curves in Figures 3 and 4 shows that the four item pools used in Test B are comparable in terms of information.

3.2 Percentages of Items in Content Categories

Tables 5 and 6 provide the percentage of items in each content category obtained based on the simulation studies for Test A and Test B, respectively. In both tables, the second column shows the mean percentages of the number of items from each of the content categories produced by the CAT simulation system, and the third column exhibits the ranges of the percentage of the number of items from each of the content categories obtained based on the simulations. For each examination, these results are identical across the four pools, as expected.

3.3 Theta Estimates

For each true theta, the mean theta estimate is computed over 1,000 replications. Figure 5 presents the relationship between the true thetas and the mean theta estimates obtained based on the simulations using each of the four pools of Test A. Figure 6 presents the relationship between the true thetas and the mean theta estimates for the four Test B pools. In both plots, the dotted vertical line indicates the cut score (-0.11 for Test A and -0.32 for Test B).

Considering Test A (see Figure 5), for each true theta value, the mean theta estimates produced by using each of the four pools are almost the same, which is seen by the overlap of the four curves. The same trend is also found for Test B (see Figure 6). For both examinations, the estimated theta values obtained using each pool are close to the true theta values except for the region near the cut score, such as -0.62 to 0.49 for Test A and -0.92 to 0.28 for Test B.

3.4 Conditional Bias (CBIAS)

Conditional biases were obtained for each of the four Test A pools, and they are shown in Figure 7. The four Test A pools have similar patterns for conditional biases. An interesting finding is that, for all the four pools, the conditional biases are positive for the true thetas between -0.10 and 0.65, whereas they are negative in the range of -0.10 to -0.925.

Figure 8 shows conditional biases for the four Test B pools. As seen from Figure 8, simulations based on the four Test B pools produce similar trends of conditional bias. In addition, the trends observed for Test A are also found for Test B. For the four pools, the conditional bias values are positive when the true theta is located in the range of -0.325 to 0.425, whereas the conditional biases are negative for the true theta values from -0.325 to -1.075. For the four Test A pools and the four Test B pools, the absolute values of the conditional biases are all smaller than 0.11.

3.5 Conditional Standard Error (CSE) of Estimated Theta

The results of two types of CSE, ACSE and ECSE, are reported in this subsection. For each item pool of Test A, the ACSEs obtained from the simulation are shown in Figure 9. The ACSEs for each Test B pool are plotted in Figure 10. As seen from Figure 9, the four Test A pools produce similar ACSEs at each true theta point, which indicates consistency of the four pools in ACSE values. As the true theta value moves closer to the cut score, the CAT system tends to administer more items (will be considered in more detail in Average Number of Items Taken at Each Theta Point section), which is associated with smaller ACSE. Therefore, the ACSEs are substantially smaller when the true thetas are near the cut score than when the true thetas are at more extreme values. The lowest ACSE is at approximately 0.17 for the Test A pools. The four Test B pools exhibit similar ACSE patterns as the four Test A pools (Figure 10). According to Figure 10, ACSEs are very similar across pools at true thetas between -1.75 to 1.55. The lowest ACSE value is approximately 0.18 for the Test B pools.

ECSEs obtained using each Test A pool are plotted in Figure 11. The general patterns of ECSE for the four Test A pools are consistent across pools, demonstrating that the four pools are comparable in terms of ECSE. Different from the ACSEs, the plots of ECSEs show a “W” shape. For the true thetas in the range approximately between -0.475 to 0.20, the ECSE increases at first, and after reaching the local maximum value (about 0.28), it declines.

From Figure 12, it can be seen that the four Test B pools tend to have similar ECSEs across the entire theta range. In addition, similar to what is found for the four Test A pools, the plots for the ECSEs for the four Test B pools also have a “W” shape.

3.6 Conditional Mean Squared Error (CMSE)

As described in the Evaluation Criteria section, two types of CMSEs were computed: including ACMSE (based on ACSE) and ECMSE (based on ECSE). Figures 13 and 14 show the ACMSEs for Test A (four pools) and Test B (four pools), respectively. From both figures, it can be seen that the four pools of the same examination have similar ACMSE patterns.

Figures 15 and 16 provide the ECMSEs for Test A and Test B, respectively. For each examination the four pools produce similar ECMSEs at each of the 81 true thetas. Different from the plots of ACMSEs, the plots of the ECMSEs for the eight pools (four pools for each examination) show a "W" shape, which is due to the "W" shape of the ECSEs.

3.7 Average Number of Items Taken at Each Theta Point

The average number of items taken by simulees at each theta point was calculated based on simulations using each pool of both Test A and Test B. The results for the four Test A pools and four Test B pools are presented in Figures 17 and 18, respectively. For both examinations, the minimum number of items administered is 60, whereas the maximum test length allowed is 250 for Test A and 180 for Test B.

As seen in Figure 17, the four lines representing the four different item pools of Test A nearly overlap completely except for some minor bumpiness over the range around the cut, which demonstrates that the four Test A pools are comparable in terms of average number of items taken at each true theta. In addition, when true theta is far away from the cut, the minimum number of items tends to be administered. However, as true theta gets closer to the cut score, more items tend to be administered to examinees and the test length is likely to reach the maximum when the true theta is very close to the cut. Numbers of items taken at each theta are similar across the four different Test B pools. Similar to Test A, the closer the true theta is to the cut, the more items are administered.

3.8 Conditional Passing Rate

Conditional passing rates obtained using the four Test A pools and the four Test B pools are plotted in Figures 19 and 20, respectively. From Figure 19, it is evident that the results of the four pools of Test A are nearly identical, suggested by the overlap of the four lines. When true theta values are less than -0.85, the conditional passing rates are all zero, meaning that none of the 1,000 examinees (here, we treat each replication as one examinee) passed the examination. When true theta is greater than 0.50, the conditional passing rates are all one, meaning that all the 1,000 examinees passed the test. For a true theta in the range of approximately -0.85 to 0.50, the conditional passing rate increases from 0 to 1 gradually.

The conditional passing rates are similar for the four Test B pools. For all four Test B pools, the conditional passing rates are 0 for the true thetas below -0.925, and they are 1 for the true thetas greater than 0.35.

3.9 Reliability and Decision Accuracy (DA)

Table 7 summarizes the reliabilities of the four Test A pools and the four Test B pools. For each item pool in both examinations, two kinds of weights were considered when computing reliability. One is the empirical weight and the other is the normal weight. For both examinations, it is evident that the reliabilities of the four pools are close to each other regardless of the type of weight used.

Conditional DAs for Test A and Test B are presented in Figures 21 and 22, respectively. According to Figure 21, for each true theta, the magnitudes of conditional DAs are very similar across the four pools of Test A. In addition, the trends of the conditional DAs are also the same across the pools. As true theta is closer to the cut, conditional DA decreases. Similar findings are obtained when considering the conditional DAs of the four Test B pools (see Figure 22).

Table 8 lists the results of overall DAs for the two examinations. As with the computation of reliability, two different types of weights were used to compute overall DAs. From Table 8, DAs are all well above 0.92 and are similar across different pools of the same examination.

3.10 Decision Consistency (DC)

As described in the Evaluation Criteria section, two types of DC were calculated including SDC and MDC. SDC was computed using a single pool, whereas MDC was computed using two different pools of the same examination.

Conditional SDCs for Test A and Test B are plotted in Figures 23 and 24, respectively. According to both figures, it can be seen that, conditional on each true theta, the four pools of the same examination have very consistent DC values. Also, the patterns of conditional DCs are very similar across the four pools of the same examination. When true theta gets closer to the cut, the conditional DC gets smaller.

Conditional MDCs for the two examinations are shown in Figures 25 and 26. For each examination, although different combinations of the four pools were used to compute the MDCs, it is clear that these different combinations led to similar MDCs conditional on each true theta. In addition, similar to what is found for conditional SDCs, smaller conditional MDC is obtained near the cut.

Table 9 provides the overall SDCs. The fourth column from the left side lists the overall SDCs for each pool of the two examinations. The last column in the table provides the average SDCs across the four pools of the same examination when different weights were used. Regardless of the weight used, all the DCs are greater than 0.90 for all the pools of both examinations. DCs are very similar across the pools for the same examination. Slightly higher DCs are found when using the empirical weights than using weights from a normal distribution.

Table 10 summarized the values of overall MDCs. The fourth column from the left side lists the overall MDCs obtained by using different combinations of the four pools of the same examination. The last column in the table provides the average MDCs across the six combinations for each examination when different weights were used. All the values in the fourth column are greater than 0.90. Different combinations of the four pools of the same examination have similar overall MDCs. Comparing SDCs and MDCs, it appears that SDCs and MDCs are, on average (the last columns in Tables 9 and 10), remarkably similar, which indicates the very high comparability of the pools.

3.11 Average Number of Items Taken and Standard Errors of Theta by Theta Intervals

As mentioned earlier, many CAT testing programs report average number of items and standard errors of theta for self-defined theta intevals. The results for the average number of items taken and standard errors of theta by theta intervals for the four Test A pools are shown in Table 11. The results for the four Test B pools are given in Table 12. For each examination, the magnitudes of the average number of items taken by theta intervals produced by the CAT simulation system are very similar across pools, and the pattern of the standard errors is highly similar across pools.

3.12 Simulation Results for Single Pool Run with Different Seeds

For the same examination, although the available item pools can be compared using the conditional and marginal statistics, the comparison cannot show what an “acceptable” difference should be. In this study, the differences that result from conducting two replications of a single pool with different random number generation seeds are used as a standard to assess the magnitude of the differences caused by using different item pools.

Figure 27 provides the results including CBIAS and ECSE obtained when a single pool (Test A the fourth pool as an example) was run twice each with a different seeds, along with the results obtained using the other three Test A pools. Differences in statistics when using a different seed are indications of random error due to using a limited number (1,000) of simulated examinees at each theta.

From Figure 27, the differences among the different pools appear to be similar to the differences caused by the different runs of the same pool. The same conclusion holds for Test B (Figure 28). This finding suggests that the differences found across pools are similar to differences on different simulation replications for the same pool, which provides additional evidence in support of the similarity of the psychometric properties of scores from the different item pools.

4 Conclusions

The present study assessed psychometric comparability of multiple item pools for two operational CAT examinations. The evaluation criteria can be used with any CAT examination. In addition, evaluation of psychometric comparability can be conducted at an early stage of test development.

For all statistics investigated there were no substantial differences across pools for the same test. Thus, the results suggest that proficiency estimates and pass-fail decisions are comparable across item pools in terms of expected scores, conditional precision of scores, and conditional passing rates for both examinations. The proficiency estimates are of similar reliability across item pools. In addition, decision consistency and decision accuracy are similar across pools. These findings support the comparability of pools as indicated by the Standards for Educational and Psychological Testing (AERA, APA, NCME, 2014), Standard 5.12, which states that "a clear rationale and supporting evidence should be provided for any claim that scores earned on different forms of a test may be used interchangeably."

One notable finding was the degree of bias in proficiency estimates for examinees with proficiencies near the cut scores (see Figures 7 and 8). These biases were not large (around .11 or less), but are clear and consistent across pools and tests. Our hypothesis is that these biases are a result of the use of the stopping rule based on standard errors.

Another notable finding is that the pattern of the analytical conditional standard errors of proficiency estimates differ from the pattern of the empirical conditional standard errors of proficiency estimates (see Figures 9 - 12). The "W" shape for the empirical conditional standard errors of proficiency estimates might be due to the use of the stopping rule based on standard errors or it might be related to the pattern of bias in the ability estimates.

The patterns of bias and standard errors should be further examined. Given the presumption that the particular patterns are closely related to the stopping rules, future research could focus on employing different stopping rules, for example: (1) a confidence interval rule with a different percentage (e.g., 68%), (2) a fixed test length (e.g., 60 or 180), (3) a varying test length with a minimum standard error, or (4) any combination of these rules. A simulation study using these various stopping rules would help better understand the results reported in the current study. It would also be interesting to investigate the impact of using different stopping rules on psychometric characteristics of the scores and fail-pass decisions.

The pools investigated in the present study contain over 1,000 items. The current large pools appear to produce scores that are psychometrically comparable, scores that are sufficiently reliable, and decisions that are sufficiently accurate and consistent. There is a possibility that smaller pools would also lead to scores and decisions that have reasonable psychometric properties. One set of future simulation studies could consider the use of different size item pools with a focus on the extent to which pools of different sizes produce comparable scores and the extent to which size of pool is related to the precision of scores and pass-fail decisions.

The present simulation study is based on a fixed cut score. Various psychometric properties considered in this study (e.g., decision accuracy) can be examined for different cut-score levels. A particular cut score could be identified that would maximize decision these psychometric properties.

5 References

AERA, APA, NCME, (2014). *Standards for educational and psychological testing.* Washington, DC: AERA, APA, NCME.

Davey, T., & Nering, M. L. (1998, September). *Controlling item exposure and maintaining item security.* Paper presented at the ETS-sponsored colloquium entitled Computer-based testing: Building the foundations for future assessments, Philadelphia PA.

Davey, T., & Thomas, L. (1996, April). *Constructing adaptive tests to parallel P&P programs.* A paper presented at the annual meeting of the American Educational Research Association, New York.

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

Owen, R. J. (1969) A Bayesian approach to tailored testing. *Res. Bull.* 69-92. Princeton, N. J.: Educational Testing Service.

Rasch, G. (1960). *Probabilistic Models for some Intelligence Tests and Attainment Tests.* Copenhagen: Danish Institute for Educational Research.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools.* (Research Report 94-5). Princeton, NJ: Educational Testing Service.

Thomasson, G. L. (1997, March). *The goal of equity within and between computerized adaptive tests and paper and pencil forms.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

van der Linden, W. J. (2006). Equating scores form adaptive to linear tests. *Applied Psychological Measurement*, 30, 493-508.

Wainer, H. (2000). *Introduction and history. In H. Wainer (Ed.), Computerized adaptive testing: A primer* (2nd ed., pp. 1-21). Mahwah, NJ: Lawrence Erlbaum Associates.

Wang, T., & Kolen, M. J. (2001). Evaluating comparability in computerized adaptive testing: Issues, criteria and an example. *Journal of Educational Measurement*, 38, 19-49.

Table 1: Layout of True Theta, Estimated Theta, and Weights

Weight of True Theta	True Theta	Theta Interval	Weight for Theta Estimate
w_1	θ_1	$\widehat{\theta}_{1,1}$	$\frac{w_1}{1000}$
		$\widehat{\theta}_{1,2}$	$\frac{w_1}{1000}$
		.	.
		.	.
		.	.
		$\widehat{\theta}_{1,1000}$	$\frac{w_1}{1000}$
w_2	θ_2	$\widehat{\theta}_{2,1}$	$\frac{w_2}{1000}$
		$\widehat{\theta}_{2,2}$	$\frac{w_2}{1000}$
		.	.
		.	.
		.	.
		$\widehat{\theta}_{2,1000}$	$\frac{w_2}{1000}$
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
.	.	.	.
w_{81}	θ_{81}	$\widehat{\theta}_{81,1}$	$\frac{w_{81}}{1000}$
		$\widehat{\theta}_{81,2}$	$\frac{w_{81}}{1000}$
		.	.
		.	.
		.	.
		$\widehat{\theta}_{81,1000}$	$\frac{w_{81}}{1000}$

Table 2: Example of Theta Intervals for Test A and Test B

Sequence of Interval	Theta Interval for Test A	Theta Interval for Test B
1	-1.91 and below	-2.12 and below
2	-1.91 to -1.51	-2.12 to -1.72
3	-1.51to -1.11	-1.72 to -1.32
4	-1.11 to -0.71	-1.32 to -0.92
5	-0.71 to -0.31	-0.92 to -0.52
6	-0.31 to 0.09	-0.52 to -0.12
7	0.09to 0.49	-0.12 to 0.28
8	0.49 to .89	0.28 to 0.68
9	0.89 to 1.29	0.68 to 1.08
10	1.29 to 1.69	1.08 to 1.48
11	Above 1.69	Above 1.48

Table 3: Summary Statistics of Item Difficulties of the Four Pools of Test A

Test A	# of Items	Min	Max	Mean	SD
	Pool 1				
Whole pool	1448	-2.2090	2.1876	-0.1916	0.9873
Content 1	279	-2.2090	2.0355	-0.1775	0.9703
Content 2	162	-2.1576	2.1768	-0.1597	0.9966
Content 3	131	-2.1704	2.1355	-0.1893	0.9712
Content 4	132	-2.1975	2.1876	-0.1884	1.0004
Content 5	132	-2.1887	2.1871	-0.2069	0.9698
Content 6	226	-2.1896	2.1608	-0.2271	0.9640
Content 7	181	-2.1153	1.9948	-0.1851	0.9811
Content 8	205	-2.2068	2.1314	-0.1962	1.0598
	Pool 2				
Whole pool	1464	-2.1963	2.1893	-0.1588	0.9813
Content 1	277	-2.0948	2.0517	-0.1491	0.9649
Content 2	163	-2.1854	2.1351	-0.1485	0.9981
Content 3	131	-2.1692	2.1297	-0.1585	0.9328
Content 4	132	-2.1875	2.0619	-0.1646	0.9838
Content 5	132	-2.1963	2.0474	-0.1744	0.9990
Content 6	236	-2.1514	2.1272	-0.1599	1.0205
Content 7	190	-2.1554	2.1893	-0.1474	0.9825
Content 8	203	-2.1787	2.0415	-0.1759	0.9759
	Pool 3				
Whole pool	1456	-2.2089	2.1876	-0.1633	0.9963
Content 1	277	-2.1969	2.1871	-0.1822	0.9888
Content 2	160	-2.1952	2.1813	-0.1454	1.0252
Content 3	130	-2.2035	2.1708	-0.1654	0.9652
Content 4	131	-2.1815	2.1876	-0.1744	0.9968
Content 5	132	-2.0509	2.1871	-0.1968	0.9760
Content 6	235	-2.1816	2.1835	-0.1374	0.9970
Content 7	186	-2.2089	2.1599	-0.1588	1.0212
Content 8	205	-1.9817	2.1684	-0.1553	1.0076
	Pool 4				
Whole pool	1448	-2.1993	2.1876	-0.1710	0.9850
Content 1	278	-2.1961	2.0253	-0.1885	0.9565
Content 2	163	-2.0727	2.1069	-0.1498	0.9799
Content 3	132	-2.1993	2.0895	-0.1872	0.9813
Content 4	131	-2.1656	2.1876	-0.1723	1.0225
Content 5	130	-2.1563	2.0501	-0.1843	0.9849
Content 6	228	-2.0794	2.1503	-0.1447	1.0033
Content 7	187	-2.0854	2.1569	-0.1578	0.9835
Content 8	199	-2.0879	2.1542	-0.1861	1.0020

Table 4: Summary Statistics of Item Difficulties of the Four Pools of Test B

Test B	# of Items	Min	Max	Mean	SD
Pool 1					
Whole pool	1230	-2.3274	2.1816	-0.2938	0.9481
Content 1	169	-2.2894	2.091	-0.3083	0.9737
Content 2	148	-2.2394	2.1591	-0.2647	0.9805
Content 3	140	-2.2782	1.8415	-0.3238	0.8975
Content 4	149	-2.3274	2.1426	-0.2647	0.9725
Content 5	161	-2.2722	2.0634	-0.3317	0.9378
Content 6	151	-2.2673	2.151	-0.3148	0.9468
Content 7	154	-2.0565	2.1294	-0.244	0.9375
Content 8	158	-1.9696	2.1816	-0.2964	0.9493
Pool 2					
Whole pool	1236	-2.3655	2.1818	-0.3012	0.9421
Content 1	168	-2.29	2.1818	-0.3019	0.955
Content 2	149	-2.0658	2.1688	-0.2691	0.9516
Content 3	141	-2.3655	2.1199	-0.2955	0.9454
Content 4	149	-1.9342	2.0539	-0.2895	0.9325
Content 5	159	-2.33	2.0879	-0.3442	0.8986
Content 6	153	-2.341	2.0655	-0.2578	0.9815
Content 7	157	-2.3468	2.1316	-0.2863	0.9522
Content 8	160	-2.3106	2.1242	-0.3595	0.9358
Pool 3					
Whole pool	1235	-2.3342	2.1866	-0.3077	0.9368
Content 1	169	-2.3308	2.091	-0.3298	0.9662
Content 2	147	-2.2986	2.1668	-0.297	0.9739
Content 3	141	-2.3342	2.0622	-0.2737	0.926
Content 4	149	-2.2845	2.1866	-0.286	0.946
Content 5	161	-2.1646	2.1024	-0.3506	0.9248
Content 6	151	-2.2941	2.1159	-0.3105	0.9289
Content 7	157	-2.2791	2.0899	-0.3054	0.9314
Content 8	160	-2.1887	2.1573	-0.3008	0.9143
Pool 4					
Whole pool	1211	-2.2894	2.228	-0.3007	0.9431
Content 1	166	-2.2894	2.1818	-0.3226	0.9639
Content 2	145	-2.1729	1.9699	-0.2699	0.9386
Content 3	139	-2.0354	1.9411	-0.2975	0.9202
Content 4	143	-2.0109	2.1426	-0.2785	0.9372
Content 5	158	-2.2722	2.0634	-0.3326	0.9487
Content 6	149	-2.2321	2.0592	-0.3024	0.9128
Content 7	152	-2.2062	2.228	-0.2519	0.9821
Content 8	159	-2.2778	2.1801	-0.3421	0.9517

Table 5: Percentages of Test Plan Categories for Test A

Content Area	Mean (Based on Simulation)	Range (Based on Simulation)
1	19.8%	18.2% - 20.0%
2	11.5%	10.3% - 11.7%
3	8.5%	8.2% - 10.0%
4	8.4%	8.1% - 9.7%
5	8.4%	7.9% - 9.5%
6	16.6%	15.3% - 16.7%
7	13.3%	12.2% - 13.8%
8	13.5%	13.3% - 14.8%

Table 6: Percentages of Test Plan Categories for Test B

Content Area	Mean (Based on Simulation)	Range (Based on Simulation)
1	15.0%	14.1% - 15.8%
2	11.6%	10.4% - 11.9%
3	10.0%	9.3% - 10.8%
4	11.6%	10.1% - 11.7%
5	13.5%	13.3% - 14.8%
6	11.7%	11.3% - 12.7%
7	13.3%	12.3% - 13.7%
8	13.4%	13.1% - 14.6%

Table 7: Summary of Reliability with Empirical and Normal Weights for Test A and Test B

Examination	Weight Type	Item Pool	Reliability
Test A	Empirical Weights	1	0.869
		2	0.859
		3	0.886
		4	0.859
	Weights from Normal Distribution	1	0.864
		2	0.851
		3	0.881
		4	0.856
Test B	Empirical Weights	1	0.870
		2	0.847
		3	0.861
		4	0.855
	Weights from Normal Distribution	1	0.867
		2	0.842
		3	0.856
		4	0.850

Table 8: Summary of Decision Accuracy with Empirical and Normal Weights for Test A and Test B

Examination	Weight Type	Item Pool	Decision Accuracy
Test A	Empirical Weights	1	0.951
		2	0.950
		3	0.937
		4	0.949
	Normal Weights	1	0.935
		2	0.931
		3	0.931
		4	0.936
Test A	Empirical Weights	1	0.941
		2	0.954
		3	0.944
		4	0.949
	Normal Weights	1	0.927
		2	0.935
		3	0.929
		4	0.932

Table 9: Summary of Decision Consistency with Empirical and Normal Weights (Single Item Pool) for Test A and Test B

Examination	Weight	Item Pool	Decision Consistency (DC)	Average DC
Test A	Empirical Weight	1	0.933	0.927
		2	0.930	
		3	0.913	
		4	0.930	
	Normal Weights	1	0.909	0.907
		2	0.904	
		3	0.903	
		4	0.910	
Test B	Empirical Weight	1	0.921	0.929
		2	0.937	
		3	0.925	
		4	0.931	
	Normal Weights	1	0.900	0.904
		2	0.910	
		3	0.901	
		4	0.906	

Table 10: Summary of Decision Consistency with Normal Weights (Across Item Pools) for Test A and Test B

Examination			Decision Consistency (DC)	Average DC
Test A	Overall Empirical Weights	Pool 1 vs Pool 2	0.926	0.927
		Pool 1 vs Pool 3	0.926	
		Pool 1 vs Pool 4	0.927	
		Pool 2 vs Pool 3	0.927	
		Pool 2 vs Pool 4	0.927	
		Pool 3 vs Pool 4	0.927	
	Overall Normal Weights	Pool 1 vs Pool 2	0.906	0.907
		Pool 1 vs Pool 3	0.906	
		Pool 1 vs Pool 4	0.907	
		Pool 2 vs Pool 3	0.906	
		Pool 2 vs Pool 4	0.907	
		Pool 3 vs Pool 4	0.907	
Test B	Overall Empirical Weights	Pool 1 vs Pool 2	0.928	0.928
		Pool 1 vs Pool 3	0.927	
		Pool 1 vs Pool 4	0.928	
		Pool 2 vs Pool 3	0.928	
		Pool 2 vs Pool 4	0.929	
		Pool 3 vs Pool 4	0.929	
	Overall Normal Weights	Pool 1 vs Pool 2	0.903	0.904
		Pool 1 vs Pool 3	0.903	
		Pool 1 vs Pool 4	0.904	
		Pool 2 vs Pool 3	0.904	
		Pool 2 vs Pool 4	0.905	
		Pool 3 vs Pool 4	0.905	

Table 11: Average Number of Items Taken and Standard Errors (SEs) of Theta by Theta Intervals for the Four Test A Item Pools

Theta Interval	Pool 1		Pool 2		Pool 3		Pool 4	
	Avg. # of Items Taken	Avg. SE of Theta	Avg. # of Items Taken	Avg. SE of Theta	Ave. # of Items Taken	Avg. SE of Theta	Ave. # of Items Taken	Avg. SE of Theta
-1.91 and below	60.0	0.267	60.0	0.268	60.0	0.271	60.0	0.268
-1.91 to -1.51	60.0	0.262	60.0	0.262	60.0	0.263	60.0	0.263
-1.51to -1.11	60.0	0.263	60.0	0.263	60.0	0.264	60.0	0.264
-1.11 to -0.71	60.0	0.264	60.0	0.264	60.0	0.264	60.0	0.263
-0.71 to -0.31	82.5	0.235	82.2	0.237	81.9	0.238	78.2	0.241
-0.31 to 0.09	250.0	0.128	250.0	0.128	250.0	0.128	250.0	0.128
0.09 to 0.49	83.0	0.237	83.5	0.237	83.9	0.237	82.8	0.238
0.49 to .89	60.0	0.265	60.0	0.265	60.0	0.265	60.0	0.265
0.89 to 1.29	60.0	0.266	60.0	0.266	60.0	0.266	60.0	0.266
1.29 to 1.69	60.0	0.266	60.0	0.266	60.0	0.265	60.0	0.266
Above 1.69	60.0	0.266	60.0	0.268	60.0	0.266	60.0	0.267

Table 12: Average Number of Items Taken and Standard Errors (SEs) of Theta by Theta Intervals for the Four Test B Item Pools

Theta Interval	Pool 1		Pool 2		Pool 3		Pool 4	
	Avg. # of Items Taken	Avg. SE of Theta	Avg. # of Items Taken	Avg. SE of Theta	Ave. # of Items Taken	Avg. SE of Theta	Ave. # of Items Taken	Avg. SE of Theta
-2.12 and below	60.0	0.276	60.0	0.278	60.0	0.275	60.0	0.277
-2.12 to -1.72	60.0	0.264	60.0	0.265	60.0	0.265	60.0	0.266
-1.72 to -1.32	60.0	0.264	60.0	0.265	60.0	0.264	60.0	0.263
-1.32 to -0.92	60.0	0.264	60.0	0.264	60.0	0.264	60.0	0.264
-0.92 to -0.52	83.3	0.231	77.0	0.24	75.2	0.244	76.2	0.24
-0.52 to -0.12	180.0	0.151	180.0	0.151	180.0	0.151	180.0	0.151
-0.12 to 0.28	76.9	0.242	75.1	0.244	77.0	0.242	77.7	0.241
0.28 to 0.68	60.0	0.266	60.0	0.265	60.0	0.265	60.0	0.265
0.68 to 1.08	60.0	0.267	60.0	0.267	60.0	0.267	60.0	0.266
1.08 to 1.48	60.0	0.268	60.0	0.267	60.0	0.268	60.0	0.268
Above 1.48	60.0	0.268	60.0	0.269	60.0	0.269	60.0	0.268

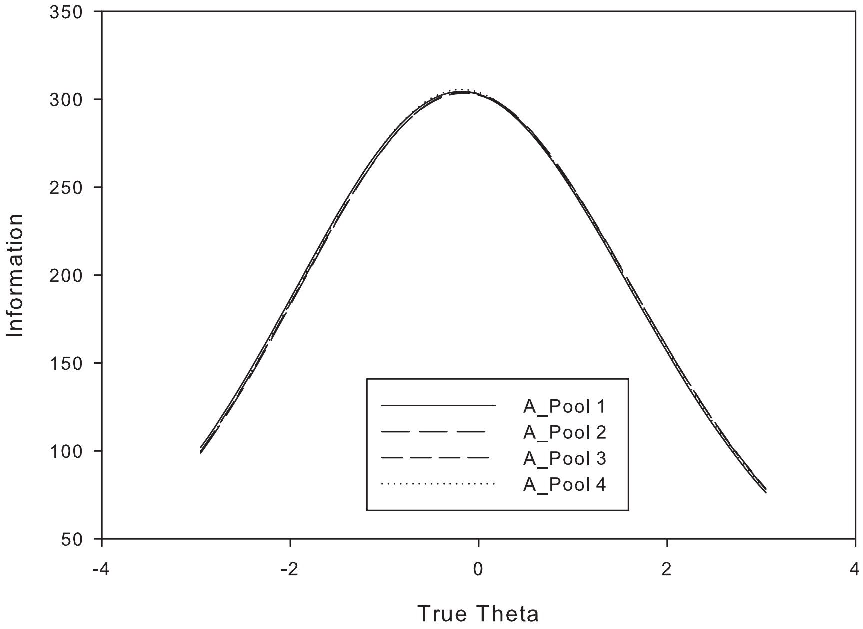


Figure 1: Comparison of pool information for Test A

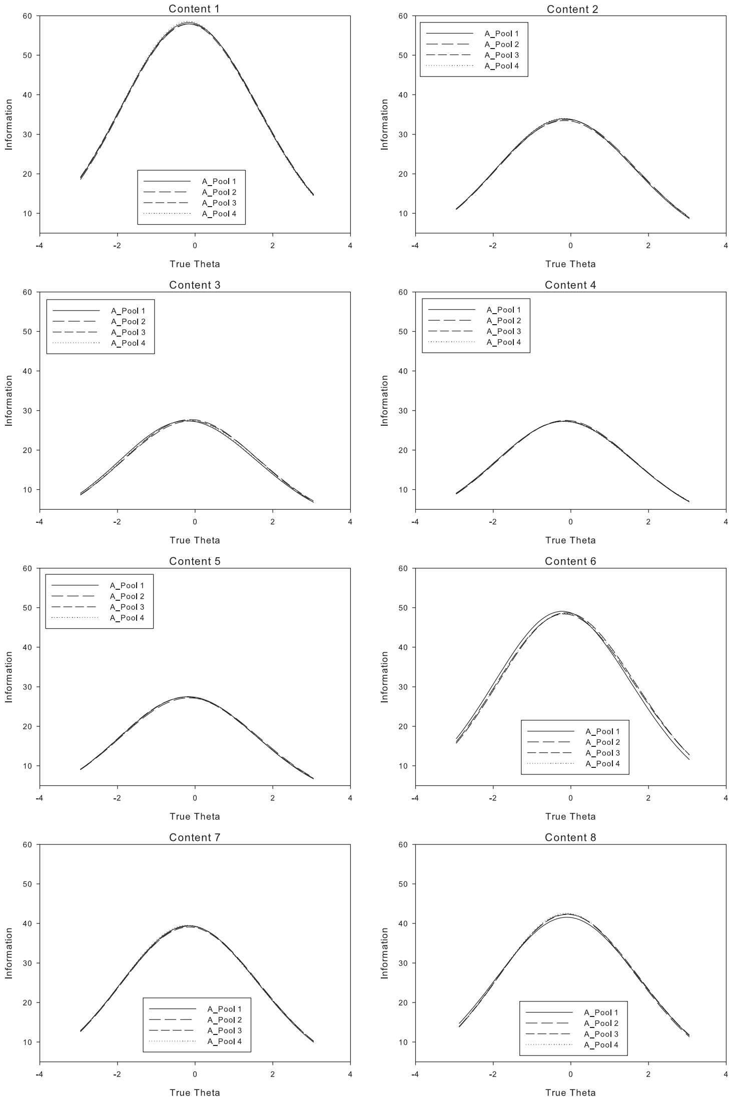


Figure 2: Comparison of information for each content category among four pools of Test A

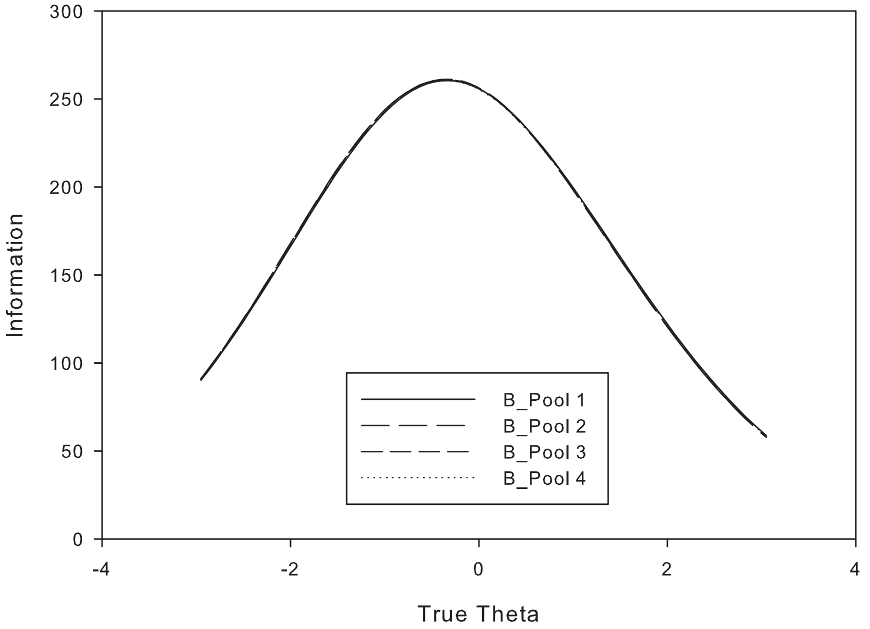


Figure 3: Comparison of pool information for Test B

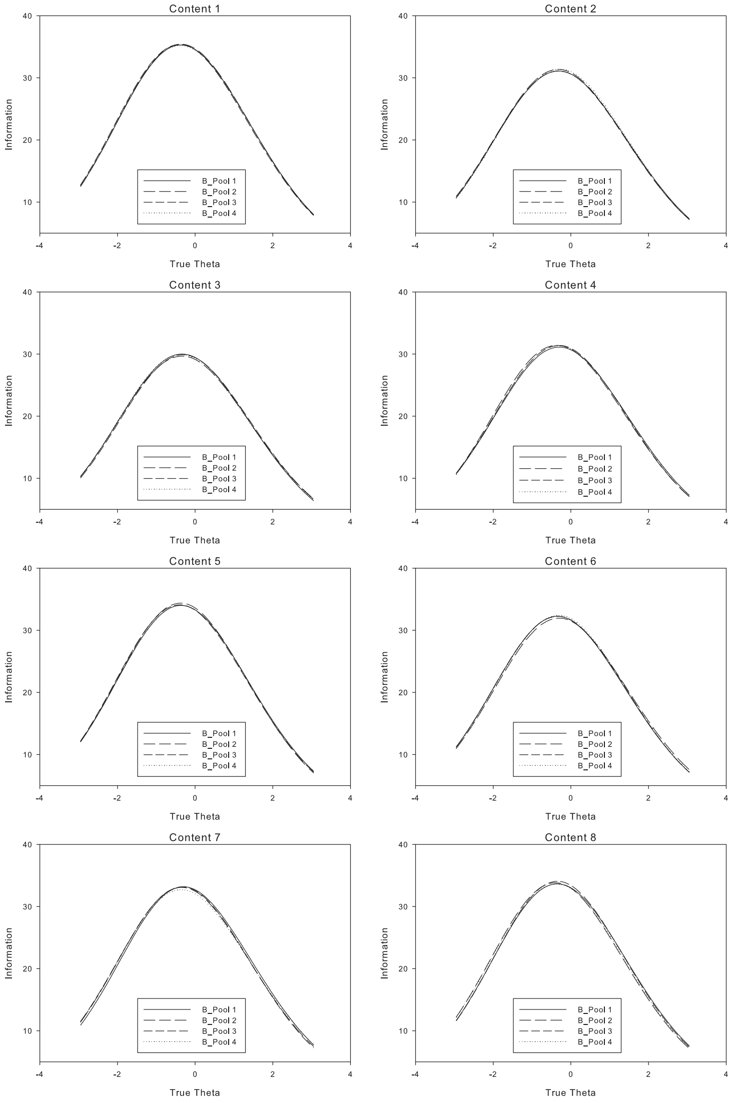


Figure 4: Comparison of information for each content category among four pools of Test B

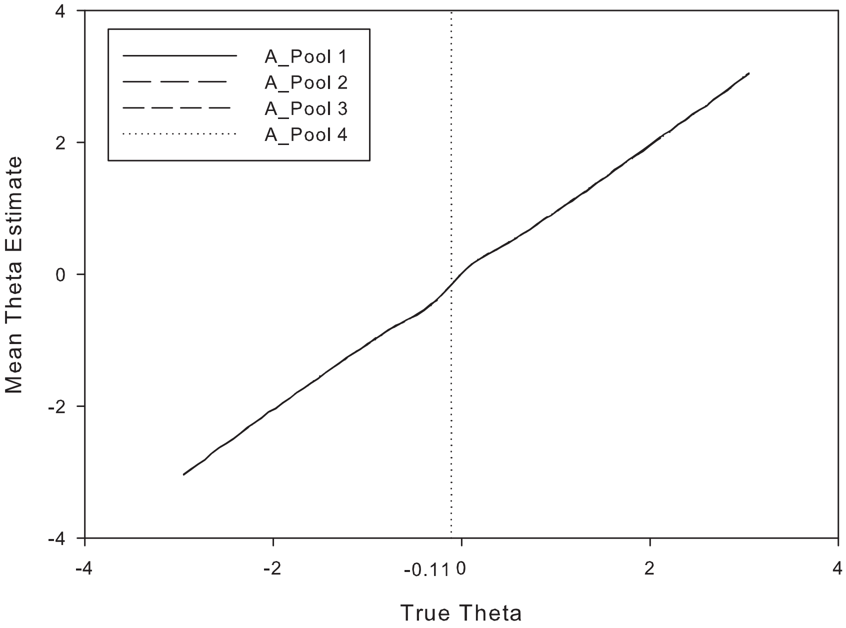


Figure 5: Mean theta estimates for each of the four Test A pools

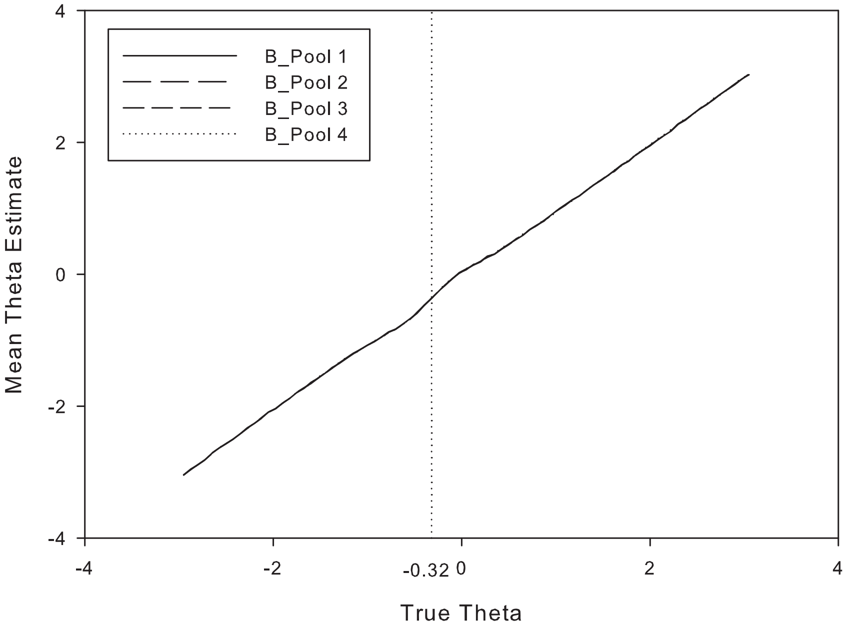


Figure 6: Mean theta estimates for each of the four Test B pools

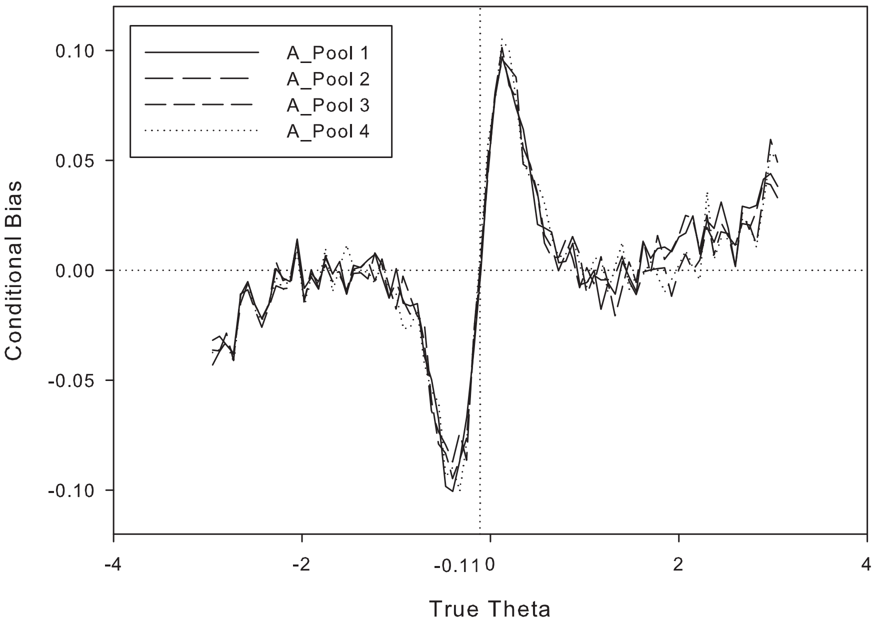


Figure 7: Conditional bias for the four Test A pools

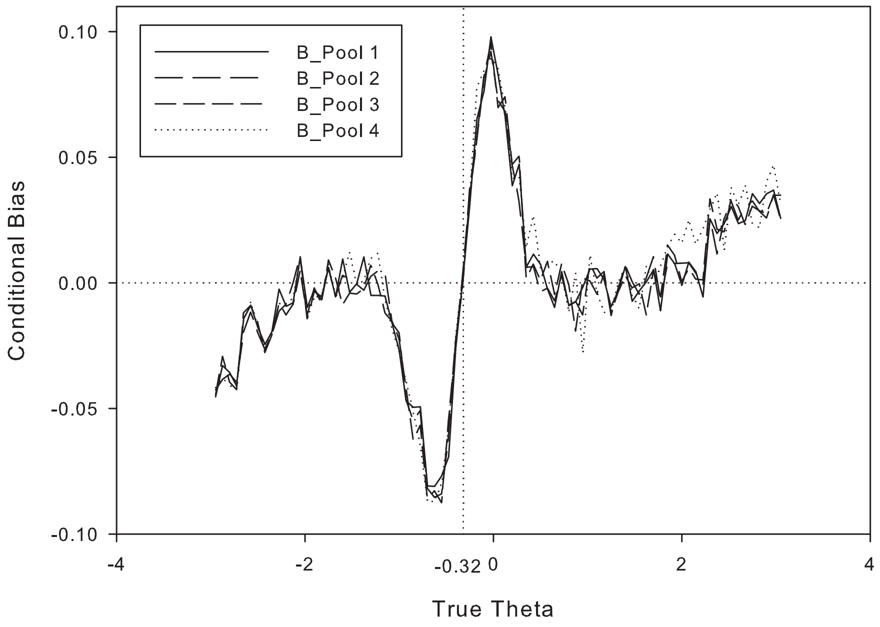


Figure 8: Conditional bias for the four Test B pools

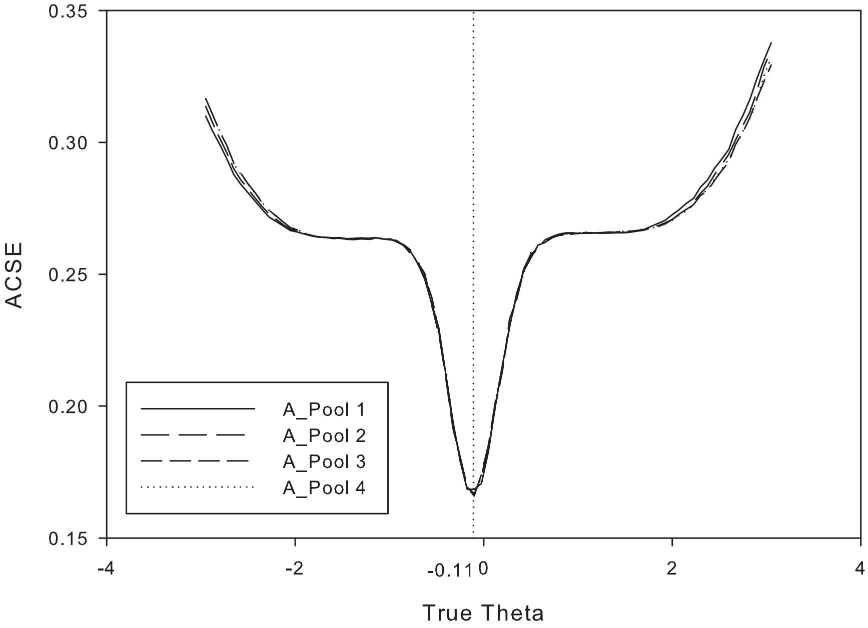


Figure 9: Analytical conditional standard error for the four pools of Test A

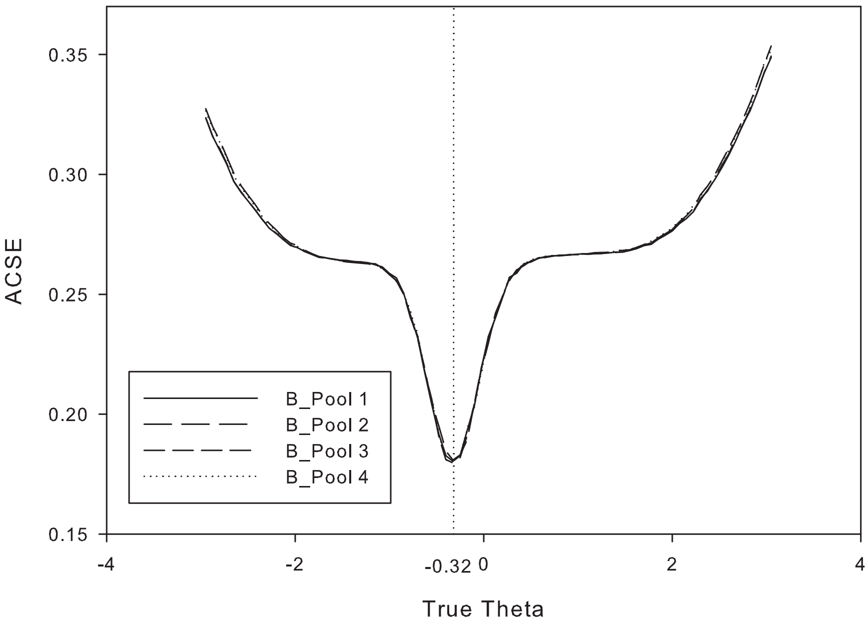


Figure 10: Analytical conditional standard error for the four pools of Test B

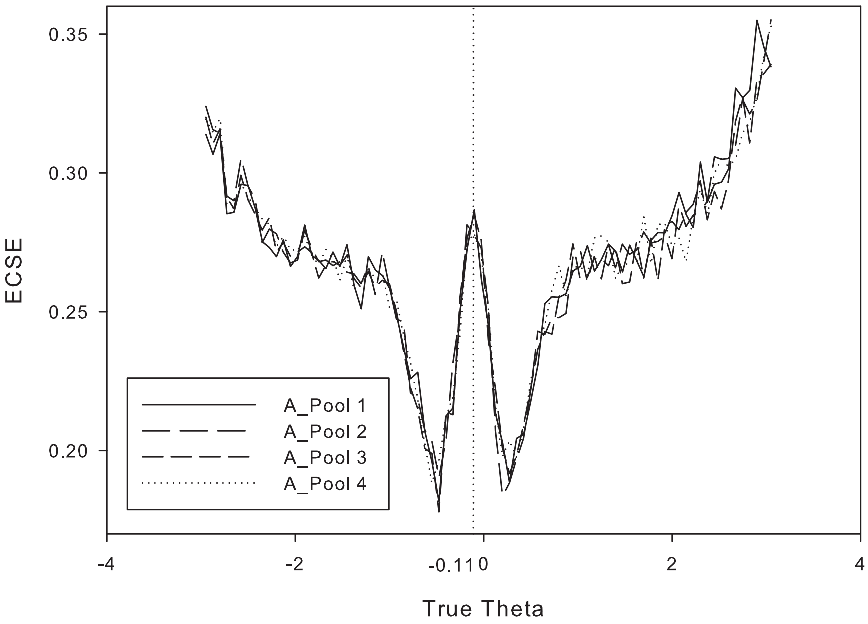


Figure 11: Empirical conditional standard error for the four Test A pools

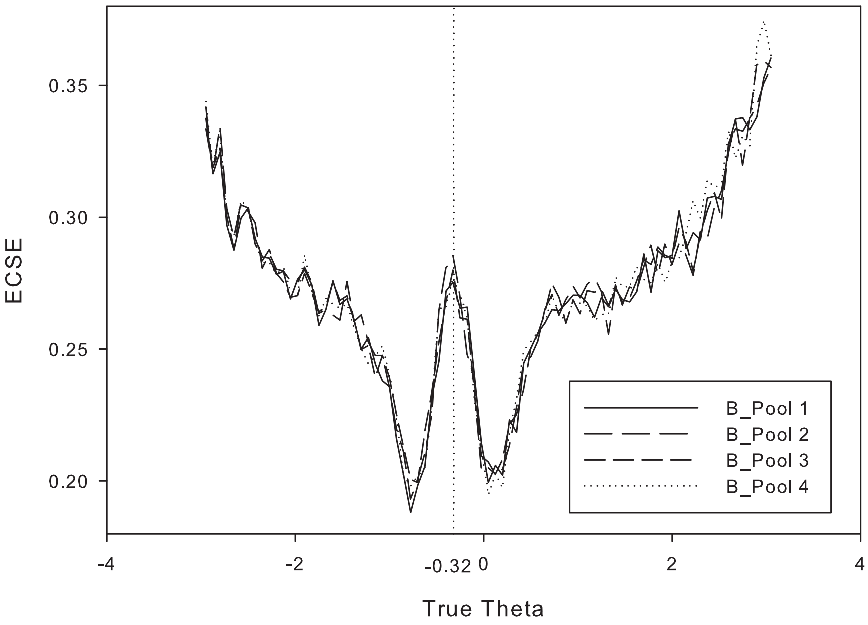


Figure 12: Empirical conditional standard error for the four Test B pools

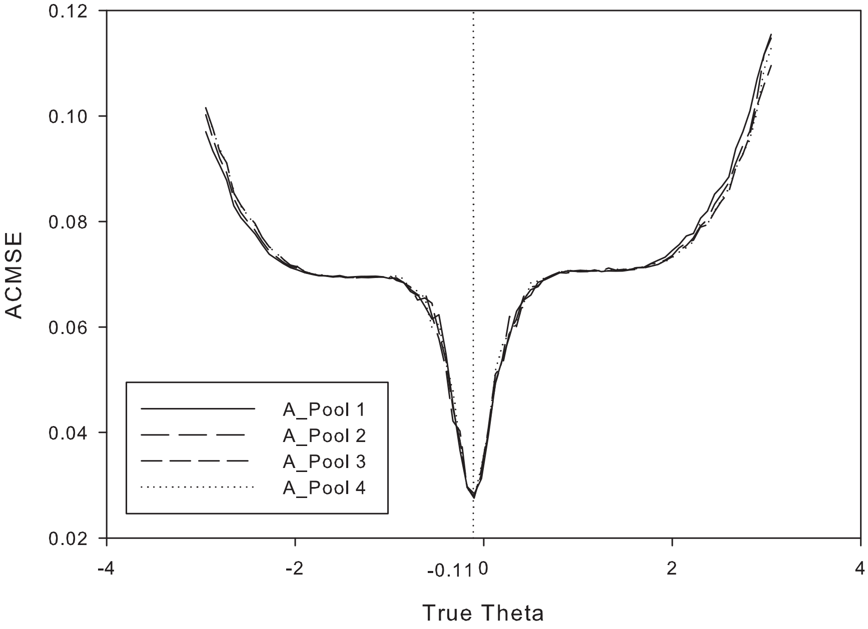


Figure 13: Analytical conditional mean squared error for the four Test A pools

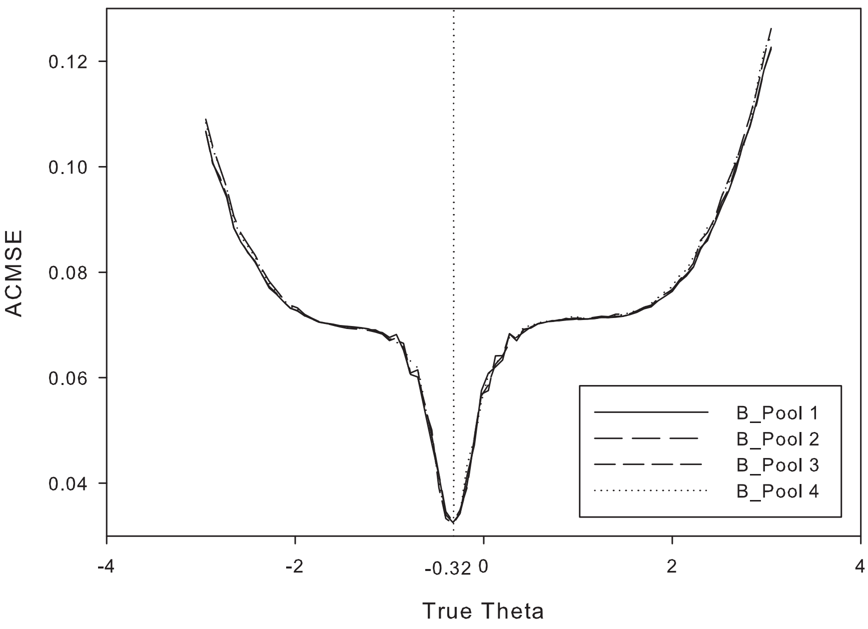


Figure 14: Analytical conditional mean squared error for the four Test B pools

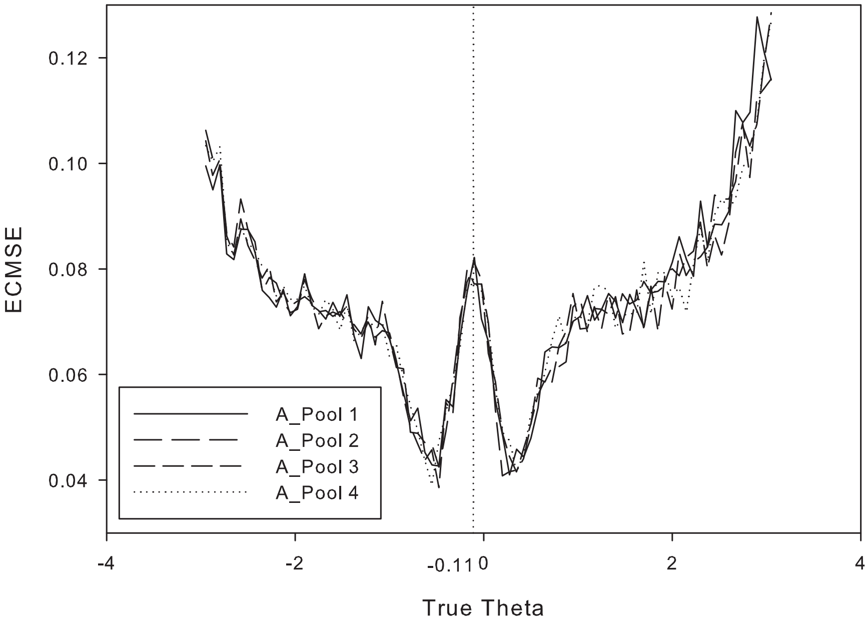


Figure 15: Empirical conditional mean square error for the four Test A pools

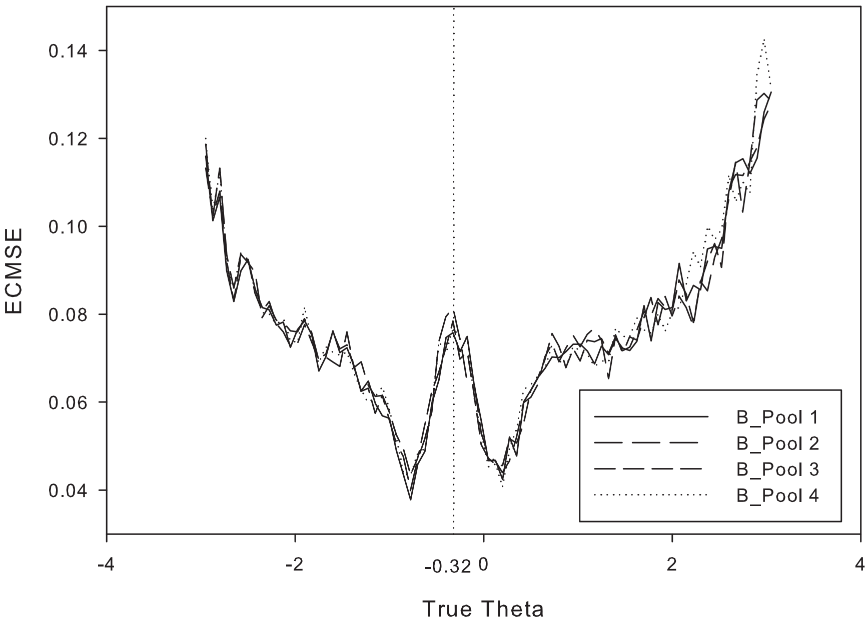


Figure 16: Empirical conditional mean square error for the four Test B pools

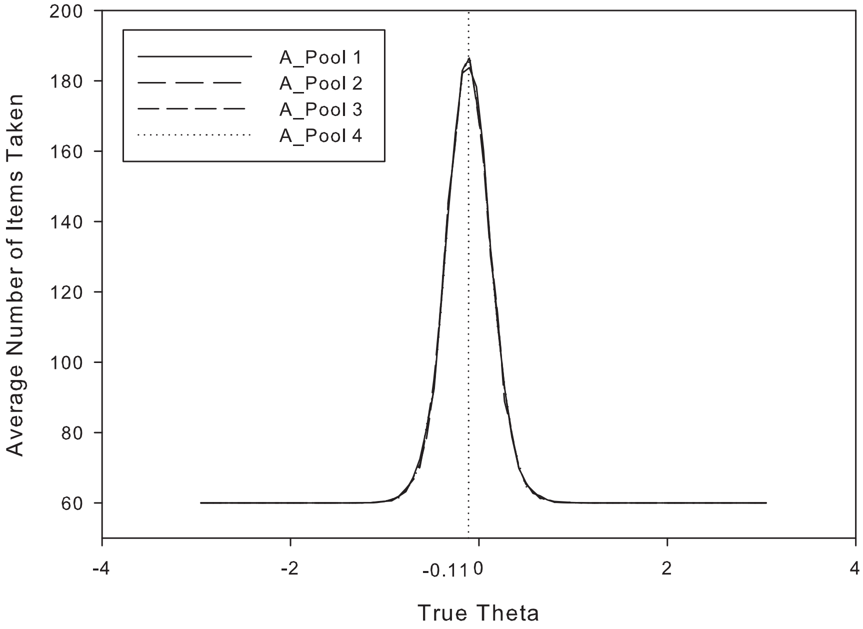


Figure 17: Average number of items taken for the four Test A pools

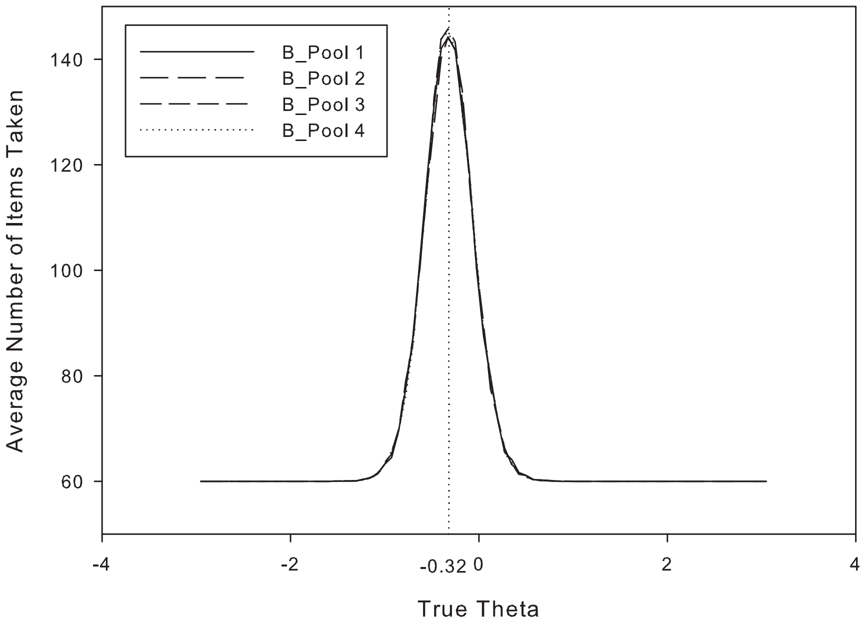


Figure 18: Average number of items taken for the four Test B pools

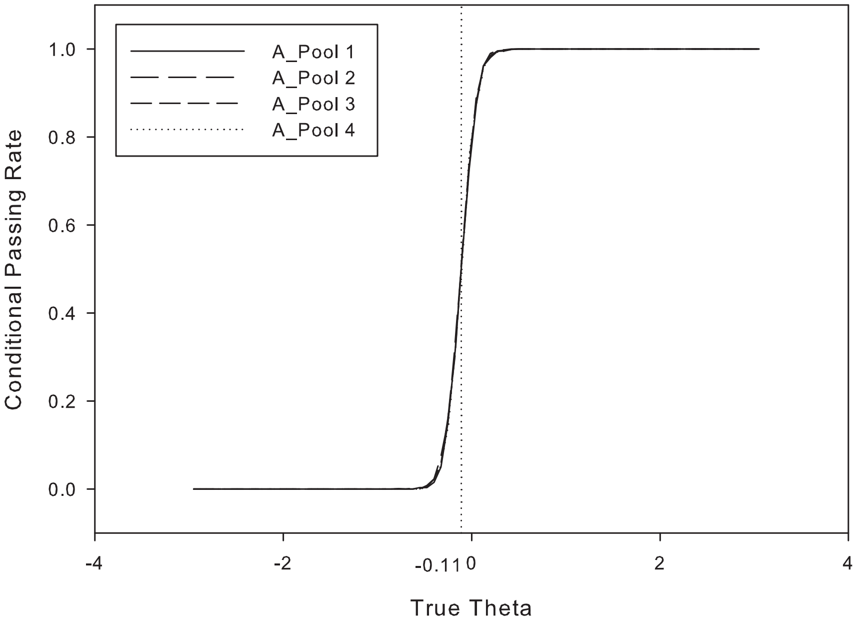


Figure 19: Conditional passing rate for the four Test A pools

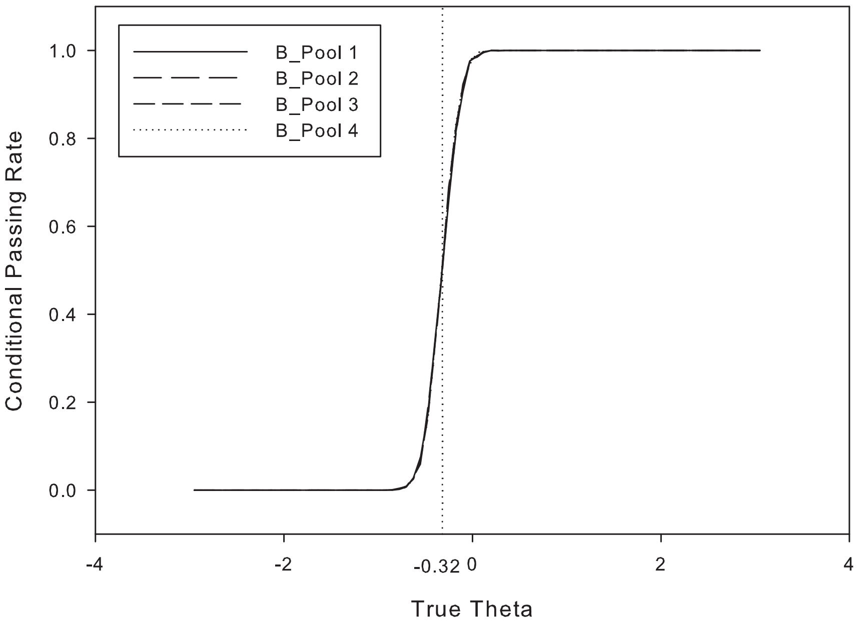


Figure 20: Conditional passing rate for the four Test B pools

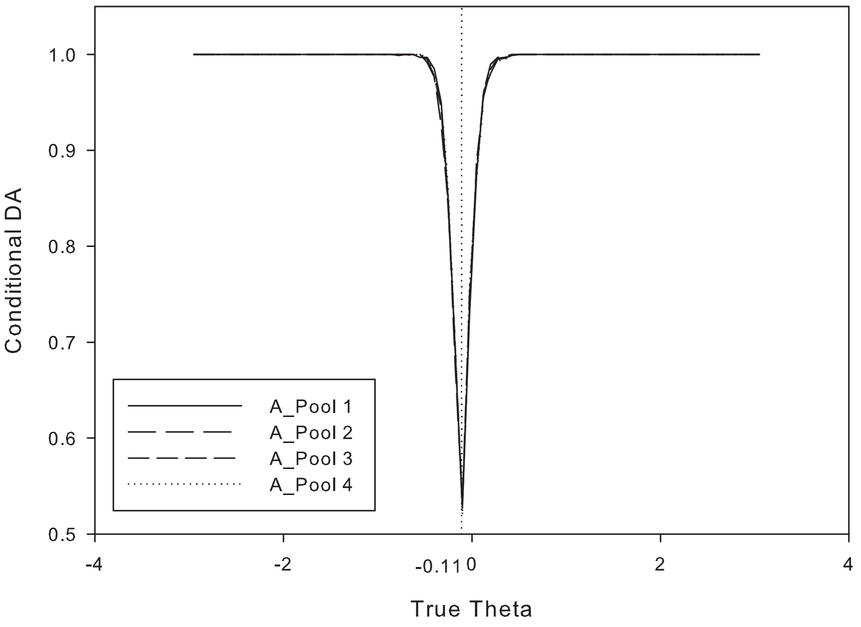


Figure 21: Conditional decision accuracy for the four Test A pools

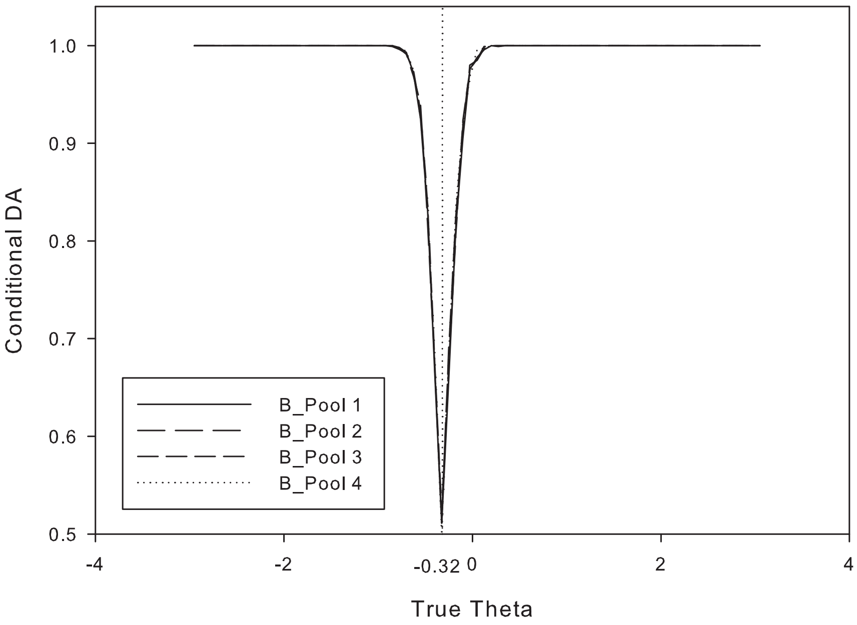


Figure 22: Conditional decision accuracy for the four Test B pools

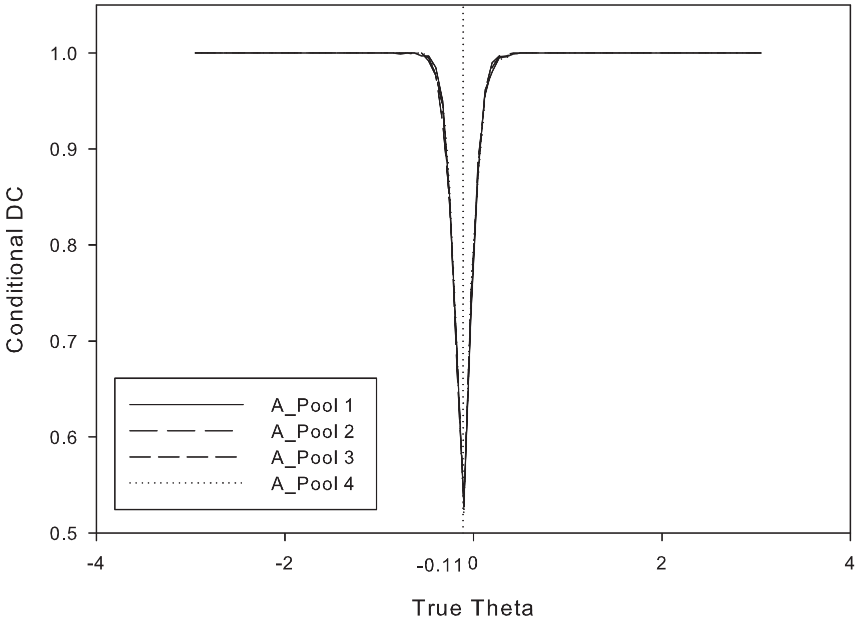


Figure 23: Conditional decision consistency based on each of the four Test A pools

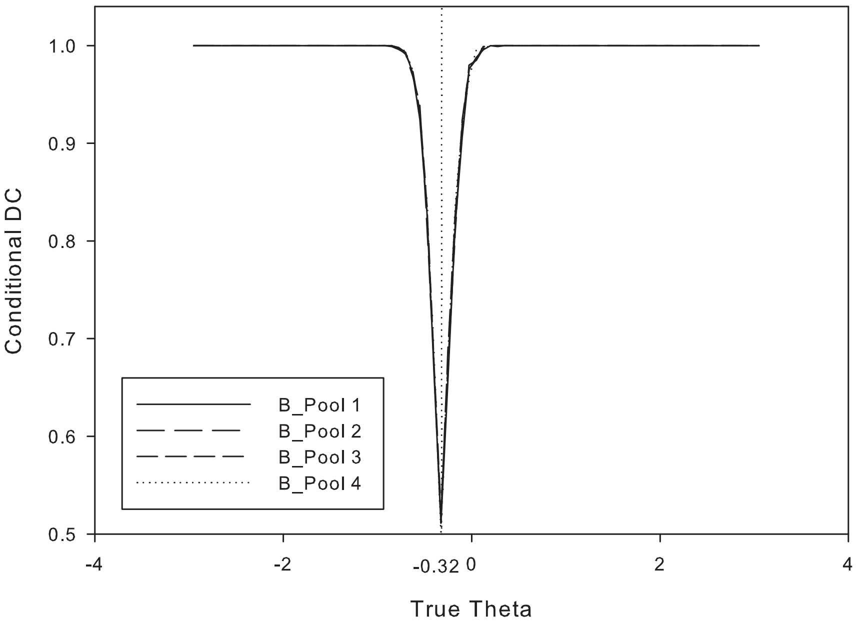


Figure 24: Conditional decision consistency based on each of the four Test B pools

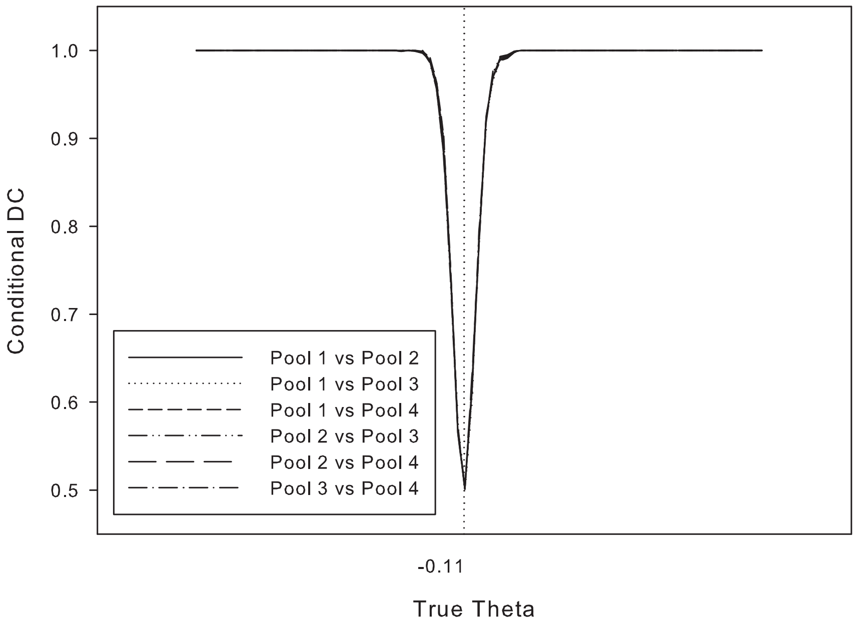


Figure 25: Conditional decision consistency based on multiple Test A pools

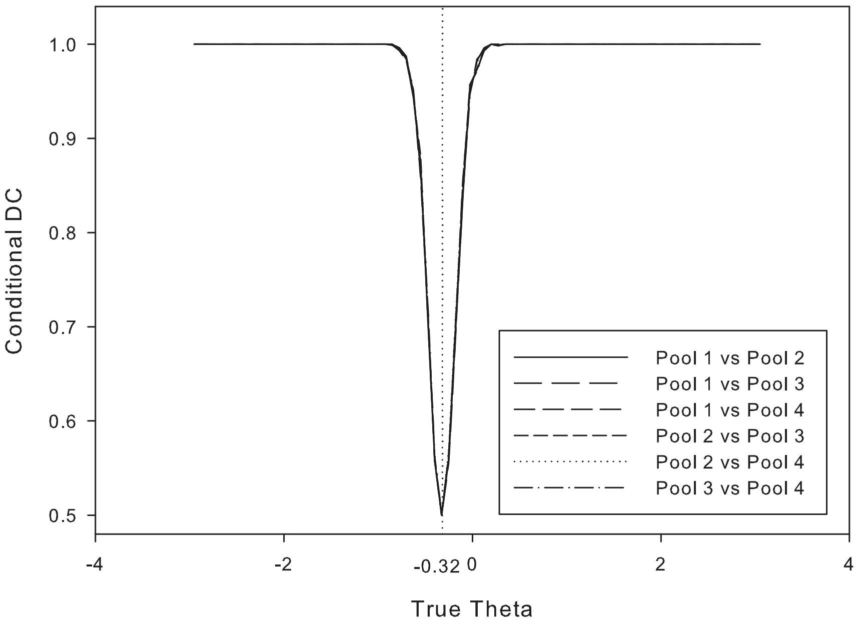
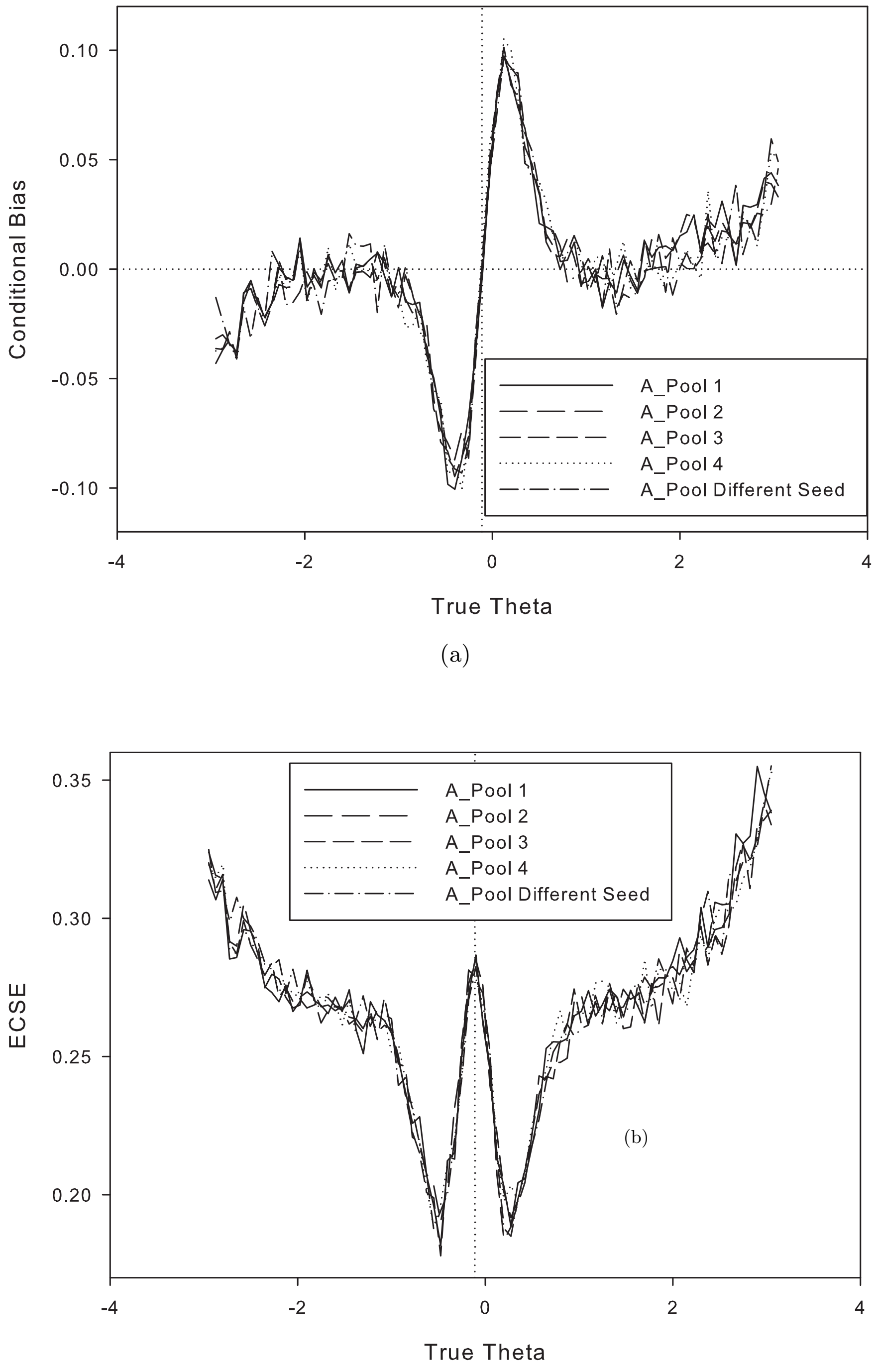


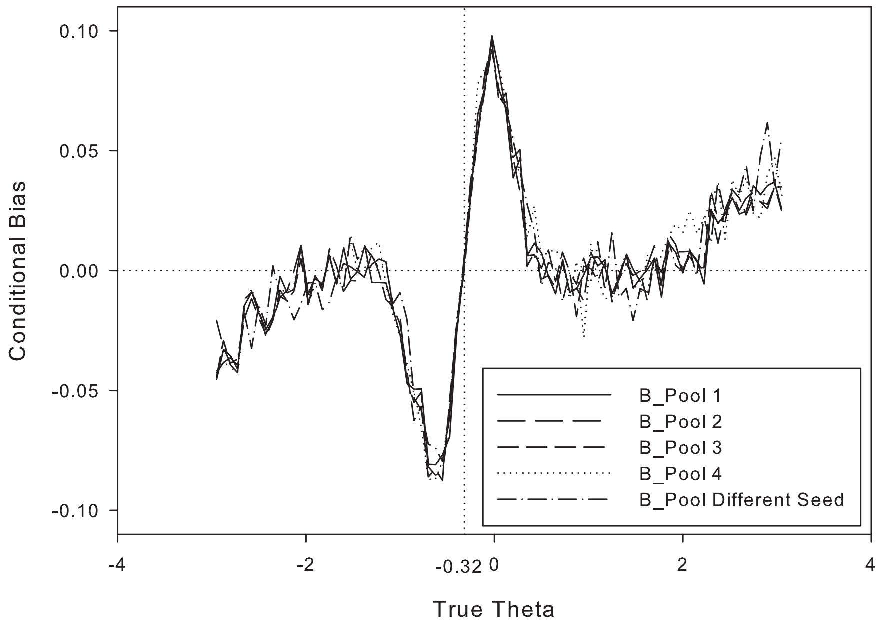
Figure 26: Conditional decision consistency based on multiple Test B pools



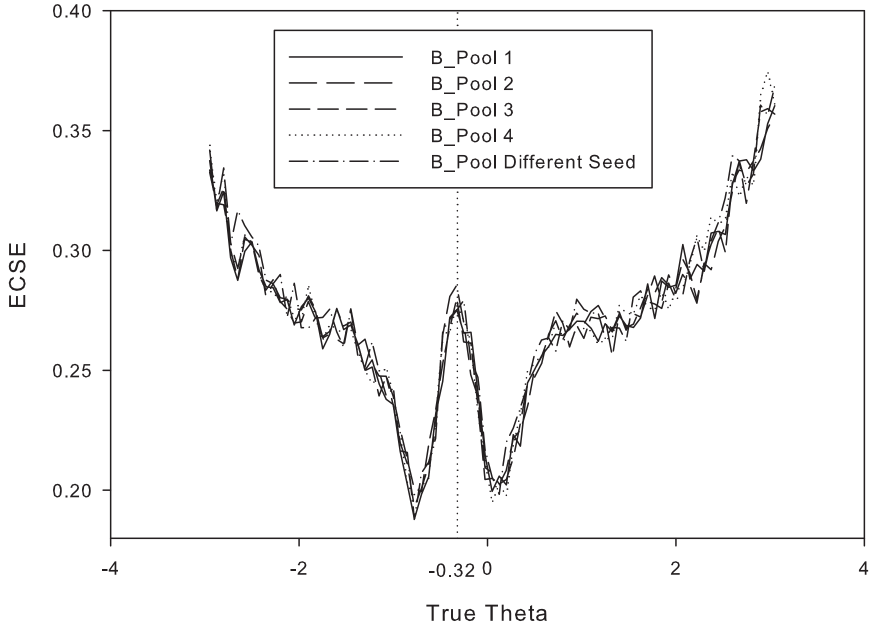
(a)

(b)

Figure 27: Comparison of simulation results with original seed and different seed for pool #4 of Test A: (a) conditional bias, and (b) empirical conditional standard error



(a)



(b)

Figure 28: Comparison of simulation results with original seed and different seed for pool #4 of Test B: (a) conditional bias, and (b) empirical conditional standard error