The Effect of Test and Examinee Characteristics on the Occurrence of Aberrant Response Patterns in a Computerized Adaptive Test [1, 2]

by

Saba Rizavi & Swaminathan Hariharan [3]
University of Massachusetts at Amherst

[1] Laboratory of Psychometric and Evaluative Research Report No. [428], School of Education, University of Massachusetts, Amherst, MA.
[2] Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA, April 10, 2001.
[3] The authors thank American Institute of Certified Public Accountants for providing funds to carry out this study. The authors also thank Professor Ronald K. Hambleton, Professor Stephen G. Sireci and Dr. Frederic Robin for their advice during the research.

Abstract

The advantages that computer adaptive testing offers over linear tests have been well documented. The Computer Adaptive Test (CAT) design is more efficient than the Linear test design as fewer items are needed to estimate an examinee's proficiency to a desired level of precision. In the ideal situation, a CAT will result in examinees answering different number of items according to the stopping rule employed. Unfortunately, the realities of testing conditions such as scheduling and improper test-taking strategies on the part of examinees have necessitated the imposition of time and minimum test length limits on CATs. Such constraints might place a burden on the CAT test taker resulting in aberrant response behaviors by some examinees. Occurrence of such response patterns result in inaccurate estimation of examinee proficiency levels. This study examined the effects of test length, time limits and their interaction with the examinee proficiency levels on the occurrence of aberrant response patterns due to rushed guessing.

The Effect of Test and Examinee Characteristics on the Occurrence of Aberrant Response Patterns in a Computerized Adaptive Test

Introduction

The advantages that computer adaptive testing offer over linear tests have been well documented. The computer adaptive test (CAT) design is more efficient than the linear test design in that with a CAT design fewer items are needed to estimate an examinee's proficiency level to a desired level of precision. This is accomplished by sequentially administering items that yield maximum precision at the examinee's current proficiency level. While this is highly desirable, a CAT with an item selection strategy that ignores such issues as content balance and exposure rates may compromise the validity of the test. Imposing content constraints and exposure controls on the CAT, while enhancing the validity of the test, imposes a considerable strain on the item pool and the administration of the CAT.

A further issue that places a burden on the CAT test-taker is the imposition of limits on time and the minimum number of items that must be attempted. In the ideal situation, a CAT will result in different examinees answering different numbers of items according to the stopping rule employed. Imposing time limits is not employed in the ideal case, since the primary objective is to estimate an examinee's proficiency level with a desired level of precision. Unfortunately, the realities of testing conditions such as scheduling and improper test-taking strategies on the part of examinees, have necessitated the imposition of time and minimum test length limits. The constraints imposed on CATs that stem from validity-related issues as well as those based on the realities of testing conditions may result in an examinee's proficiency level being estimated incorrectly.

Problems with estimating an examinee's proficiency level may occur for several reasons. An examinee may exhibit an aberrant response pattern such as responding correctly to a "difficult" item and incorrectly to an easy item. When such response patterns occur, especially with a three-parameter item response model, the likelihood function will not have a proper maximum, resulting in an inadmissible estimate for the examinee's proficiency level. Another problem in estimating examinee proficiency is that the item response model employed in CAT may not adequately model the examinee's performance. The existence of aberrant response patterns and the attendant inadmissible proficiency level estimates have been discovered in several testing programs.

This area of research has been known as appropriateness measurement in the past (Yi & Neiring, 1999; Drasgow & Levine, 1986) and as person fit analysis more recently (Meijer & Neiring, 1995; Reise & Due, 1991). A variety of person-fit indices have been proposed to detect such aberrance and a great deal of research has been devoted to this issue (Bracey & Rudner, 1992; Kogut, 1987; Kogut, 1986). While a variety of research has been conducted on detecting the extent of aberrance in a CAT, hardly any studies looked at the effect of test or examinee characteristics on the response aberrance and how that aberrance in turn reflects in the ability estimates.

Purpose

The main purpose of the current study was to examine the reasons behind the occurrence of aberrant response patterns and their effect on the estimation of the proficiency level of an examinee. (It should be noted that this study was not aimed at examining procedures for detecting aberrant response patterns. Numerous procedures have been

developed for detecting aberrant response patterns using appropriateness measurement indices and fit indices). In order to explore the issue of aberrant response patterns, this study examined the effect of test length and time constraints on the occurrence of aberrant response patterns. The effect of interaction between the examinee proficiency level and various test lengths and time constraints was also studied. The study also examined the occurrence of such patterns in a variable length computer adaptive test designed for classification purposes.

An important factor in CAT is that of content constraints. The purpose of imposing content constraints is to enhance the validity of the test by ensuring that the content domain is represented. In fact, it can be argued that guessing on a relatively easy item by an examinee with high ability may be the result of content constraints; examinees with high ability value may not know a particular area of content and hence guess on an easy item from this content area. This will result in an aberrant response pattern. Thus, the existence of content constraints may provide an explanation of why aberrant response patterns occur.

While content constraints were not explicitly imposed in the study, the effects of content constraints can be studied to some degree from the proposed design. Since the effect of examinees guessing at various points on their response patterns can be interpreted from the content-constraint perspective, the net effect of imposing content constraints can be examined.

The study looked at the issue of aberrance for fixed length achievement tests and both fixed and variable length mastery tests. Hence, the effect of test and examinee characteristics on the response aberrance was studied in the context of estimation accuracy as well as decision accuracy for adaptive tests designed and administered for a specific purpose. The research also considered the effect of aberrance on pool utilization and vice versa.

Aberrant Response Patterns

Statistical Definition of an Aberrant Response in IRT framework

The lower the probability of the response determined by the IRT model parameters, the more aberrant the response (Reise & Due, 1991). In general, the aberrance is statistically defined in terms of the Maximum likelihood function as the value of the function decreases due to the occurrence of an unlikely response, given the model. For example, a pattern of correct guessing on a set of difficult items by a low ability examinee will adversely affect the Likelihood function, which in turn results in an inaccurate estimate of the final ability of the examinee.

Definition of Aberrance in Terms of Information

An aberrant response is the one that provides less psychometric information (Lord, 1980) for estimating ability than would be expected by the parameters of a specified IRT model. Here the aberrance is defined in terms of the test information function, as it's value decreases if aberrant responses occur.

Appropriate Measurement or Person-fit Research

Since the occurrence of aberrant or non-model fitting responses for examinees frequently results in incorrect score reporting, the whole purpose of a test is hence defied. Over the past 25 years, this area received a lot of attention where researchers have tried to detect examinees with such non-model fitting response patterns or in other words "misfitting persons" or "inappropriate" score or response patterns. As mentioned before, this area of research has been known as appropriateness measurement in the past (Yi & Neiring, 1999; Drasgow & Levine, 1986) and as person fit more recently (Meijer & Neiring, 1995; Reise &

Due, 1991). While almost all of the studies in this area have been conducted to address the issues of detection, they do provide us with an idea of the kind of response aberrance that could be expected and specifically the kinds of simulation studies that would be suitable to various situations. Readers that are interested in a detailed overview of and recent developments in the person-fit detection methods, refer to Meijer and Sijtsma (1994).

Mis-fitting or Aberrant Response Patterns

An item response model can be inappropriate for an examinee even though the model may be appropriate for the whole group of examinees. The model may be inappropriate or the responses may be aberrant for a number of reasons. Researchers have observed that in a paper and pencil testing situation, for example, examinees may skip an answer on the test without skipping the item on the answer sheet. In some cases, they might turn easy items into "tricky" hard questions thus creating difficulty in the items that was not in the test design (Mcleod & Lewis, 1996). For the remainder of the test, the IRT model falsely assumes that the examinees are answering the items based on their true abilities thus resulting in low scores for such examinees.

Another situation arises, when examinees cheat or copy some answers from the other test takers. The ability estimate in this case will depend on the other test takers' abilities and may result in unexpectedly high scores if the other test taker is a high ability examinee. The inappropriateness of a response according to a model, might also be due to the violations of the underlying assumptions such as invariant ability over items/subtests, unidimensionality or local independence assumption in case of most commonly used IRT models (Glas & Meijer, 1998). Guessing, cheating, memorization, creativity, fumbling, and fatigue, for example,

result in the violation of the assumption of invariant ability across items. Cheating and memorization also results in the violation of the assumption of unidimensionality.

The issues are more serious in case of computer adaptive tests that bring along with them numerous allowances but also constraints such as the prohibition of item review or item omits. The examinees therefore, intentionally or unintentionally come up with innovative techniques to beat the test. In a CAT, for example, in addition to the forth-mentioned behaviors, the issue of memorization can be more serious compared to paper and pencil tests. Research shows that if the item pool is smaller, the examinees might inflate their scores by memorizing difficult items and using those items to route themselves to more memorized items (Mcleod & Lewis). While a number of aberrant behaviors may occur in a CAT, typical forms of aberrant response behavior are guessing and cheating which may result in spuriously high or low scores (Glas & Meijer, 1998).

Design

In this study, four different testing scenarios were examined; fixed length performance tests with and without content constraints, fixed length mastery tests and variable length mastery tests without content constraints. For each of these testing scenarios, the effect of two test lengths, five different timing conditions and the interaction between these factors with three ability levels on ability estimation were examined. For performance tests, the lack of items in a certain content area was simulated to look at the effect of their interaction with aberrance. For fixed and variable length mastery tests, decision accuracy was also looked at in addition to the estimation accuracy.

The interaction of aberrance with the total pool information was studied briefly during the course of the research; however, the results of that particular analysis are not central to the study.

The time limits were imposed by the introduction of random guessing after the examinee had answered a certain percentage of items out of the total test length. The response patterns were simulated; first by simulating the item and ability parameters and then using the item parameters obtained from a high stakes test results.

In the first step of the study, an item pool of 600 items was established using simulated item parameters. A fully adaptive test was administered to 1200 simulees (100 examinees at 12 ability levels) on 30 and 75 item fixed length tests using CBTS (Robin, 2000). For each test, every item was taken as a multiple-choice item with 4 alternatives. The CAT program used Weighted Deviations Model for item selection (Stocking & Swanson, 1993, Stocking, 1996) and Stochastic Item Exposure control methodology to control for item exposure (Reveulta & Ponsoda, 1998; and Robin, 1999). This administration of the CAT was called the CAT delivered under "Null" conditions. The "experimental" conditions included simulations conducted to generate data depicting response aberrance. To simulate random guessing at certain points in the test, the probability of a correct response was changed to the chance probability. Such conceptual framework for simulating random guessing behavior has been frequently used for detecting such aberrance using IRT based person-fit indices (Kogut, 1986; Meijer, 1994). The chance probability in this case was 0.25 because of the four alternatives to each item. In the experimental conditions, guessing was introduced after a certain percentage of items had been delivered. The following formula was used:

Number of items (n) = factor x total number of items/100 (1)

where factor is the percentage of items. Hence the regular administration of the CAT continued till the algorithm hit equation 1. At this time, the probability of the correct response was changed to:

$P(\theta)$ = 1/number of alternatives (2)

The following logic applies: For a fully computer adaptive test, an item is selected based on the information that items in the pool carry at the provisional ability estimate and administered after other constraints have been applied. The ability estimate is computed using Maximum Likelihood estimation (Hambleton, Swaminathan & Rogers, 1991) based upon the responses to the previously delivered items. If responses to the first n items are generated according to the item response model (3PL in our case), the random guess at the $(n+1)^{th}$ item disrupts the selection algorithm. The ability estimate that was expected for that particular examinee on an item with a particular item difficulty will not be produced. In fact, a wrong provisional ability estimate will be computed unless the guessed answer is the same as the expected answer. The information function will then be calculated for the remaining items in the pool. Hence the $(n+2)^{th}$ item will be the one that provides maximum information at an incorrect ability estimate, thus resulting in an item which is not well targeted to the person's true ability. Since, we assume that the random guessing will continue till the end, the same process will be repeated again and again thus resulting in a final ability estimate much different from the true estimate.

The experimental conditions were thus replicated by changing the factor to 90%, 75%, 50%, and 25% of the items for two different test lengths of 30 and 75 items.

Next, varying the proportion of examinees that guessed after a certain percentage of items had been administered resulted in another set of simulations. For example, out of 60%

of examinees that were flagged to start random guessing in a CAT administration, 80% started guessing after 90% of the items had been administered while 20% of the group started guessing after 75% of the items had been administered. Those percentages were then manipulated to represent other patterns of guessing behavior at each ability level.

The results from the previous steps of the study were then used to examine and describe the response patterns and their effects on the CAT administration in the following steps of the study.

Next, a similar experimental design was replicated. However, this time the item pool was calibrated using parameters from the November 1996 to 1998 administrations of the American Institute of Certified Public Accountants licensure examination. CATs were simulated for fixed length performance testing with and without content constraints.

In the next step of the study, an adaptive mastery test was simulated where the points after which examinees' guessed were the same as performance testing. Since random guessing is expected to have a significant impact on mastery decisions for people with abilities close to the cut scores, this analysis was particularly useful. The cut-scores approximately similar to AICPA cut-scores were used for the study and the classification or mastery decision was defined as master/pass or non-master/fail.

The final step of the study involved simulating responses for variable length adaptive classification tests where the test for a given examinee depended on a certain stopping criterion. In this case, testing stopped when a required confidence level had been attained in a pass/fail decision.

Item Pool Characteristics Using Simulated Parameters

The ability parameters of the examinees that were meant to take the adaptive tests were drawn from a normal distribution with mean of 0.0 and standard deviation of 1.0. The following table depicts the specifications that were used to generate item parameters:

Table 1: Item Parameter Distribution

	Distribution	Minimum	Maximum	Mean	Std. Dev.
A	Log-Normal	0.50	1.60	0.80	0.20
B	Normal	-2.50	2.50	0.00	1.20
C	Log-Normal	0.00	0.50	0.15	0.10

Item Pool Characteristics Using AICPA Parameters (without Constraints)

In order to look at the issue of aberrance in CAT using AICPA item parameters, a careful selection of items is necessary to create a representative item pool. For similar reasons, the data from November 1996 to 1998 AICPA administrations were used. November results were used because of the similarity in the ability distributions that took the test at a particular time of the year. The November administrations were selected instead of May administrations of the tests as for November administrations, results from three administrations were available to us. In other words May data were available only for 1997 and 1998 administrations of the test. Two tests with similar number of items and rather unique content were chosen for the purpose of our analyses. The two tests were Audit and Accounting and Reporting (ARE). Although analyses were performed on both tests, results from Audit will be explained thoroughly in this study while the results from ARE will be included in appendix D for readers' interest. To look at the distributions, following steps were carried out:

a. Multiple Choice data for each administration of the two tests were cleaned for missing cases. The multiple-choice section for each test was composed of 75 items.

b. Computer program BILOG was then used to calibrate the tests. Normal priors were set on the threshold parameter for better estimation.

c. The ability estimates from phase 3 of each of the six response matrices (3 administrations x 2 tests) were read into SPSS to look at the ability distributions.

d. Histograms of the six ability distributions were plotted. For each administration, the ability distribution was approximately normal with a mean of 0 and standard deviation 1. These distributions are depicted in figures (1) to (6) in appendix A.

Table 2: Ability Parameter Statistics

Descriptive Statistics	**Audit**			**ARE**		
	1996	**1997**	**1998**	**1996**	**1997**	**1998**
N	50317	52292	48699	50448	52799	50554
Mean	0.0	0.0	0.0	0.0	0.0	0.0
Stdev	1.0	1.0	1.0	1.0	1.0	1.0

As shown in figures, a very small percentage of examinees had ability levels in the tails of the distribution; majority of examinees were concentrated in the range of –3 and +3. Hence the hypothesis of the similarity of the examinee ability distribution for the November administrations was supported by the analyses. Next, a number of steps were performed to create a representative item pool using AICPA item parameters.

Items with difficulty parameters greater than 3.0 and less than –3.0 as well as items with item discrimination greater than 2.0 were deleted. Such items do not contribute in providing information about the examinees and hence were not included in the pool. The ability distributions for all November administrations had a mean of 0 and standard deviation of 1; equating of item parameters was not deemed necessary for our particular analyses. Item parameters from the several administrations were, therefore, combined to create a representative item pool. For audit, after combining the three administrations, 223 items were available to us in the pool, while 206 items were available for the ARE. In order to create an item pool that could be sufficient for a CAT with 75 items, 600 items was considered as a "sufficient" pool size. Hence, item parameters for the available items were examined to clone the remaining items in the pool.

Histograms as well as P-P probability plots of the existing item parameter distributions were carefully analyzed to simulate the remaining items. The P-P chart plots a variable's cumulative proportions against the cumulative proportions of a number of test distributions. Probability plots are used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line. The probability plots were analyzed for various matching distributions to decide on the most closely matched distribution. The descriptive statistics of the selected distributions are shown in table 3 and the actual P-P plots for those distributions are presented in figures (7) thru (12) in appendix A. Computer program CBTS was then used to generate the remaining items. A representative item pool of 600 AICPA items parameters was now available to us.

Table 3: Item Parameter Statistics for Audit and ARE

	AUDIT				ARE			
	Minimum	Maximum	Mean	Std. Dev.	Minimum	Maximum	Mean	Std. Dev.
a	0.24	1.65	0.78	0.29	0.19	1.72	0.78	0.33
b	-2.87	2.79	0.01	0.94	-2.89	2.96	0.34	1.12
c	0.07	0.46	0.24	0.07	0.08	0.50	0.25	0.09

Item Pool Characteristics Using AICPA Parameters with Content Constraints

The AICPA examinee booklet provides a detailed outline of the major content areas divided into several sub-areas. Audit, for example, consists of four major content areas that are in turn divided into sub-areas that range from 3 to 13 in numbers; some of these sub-areas are then refined into finer content strands. The ARE test is composed of six major content areas and each content area is then subdivided into finer strands. The examinee booklet also lists the percentage of items that are drawn from each content area while constructing the test.

For our analyses, we used the major content categories and the resulting CAT contained similar proportions of items as represented in P&P version of AICPA. These content areas and the respective percentages of items represented in the test are shown in the following table:

Table 4: Content Specifications for Audit

Content	Topic	%
1	Plan the engagement, evaluate the prospective client and engagement, decide whether to accept or continue the client/engagement and enter an agreement	40
2	Obtain and document information to form a basis for conclusions	35
3	Review the engagement to provide reasonable assurance that objectives are achieved and evaluate information obtained to reach and to document engagement conclusions	5
4	Prepare communications to satisfy engagement objectives	20

Table 5: Content Specifications for Accounting and Reporting

Content	Topic	%
1	Federal taxation --- individuals	20
2	Federal taxation --- corporations	20
3	Federal taxation --- partnerships	10
4	Federal taxation --- estates and trusts, exempt organizations, and preparers' responsibilities	10
5	Accounting for governmental and not-for-profit organizations	30
6	Managerial accounting	10

Both item pool and the test were therefore constructed to represent the respective proportions of various content categories. The number of items used for the formation of pool and the test are given below (for both tests, pool size=600; test length=75, 30):

Table 6: Pool and Test Content Composition

Content	Number of items (Audit)			Number of items (ARE)		
	Pool	Test (75)	Test (30)	Pool	Test (75)	Test (30)
1	240	30	12	120	15	6
2	210	26-27	10-11	120	15	6
3	30	3-4	1-2	60	7-8	3
4	120	15	6	60	7-8	3
5				180	22-23	9
6				60	7-8	3

The simulations were then performed for CAT with content constraints and compared to the scenarios where such constraints were not included.

Mastery testing using AICPA parameters

As mentioned earlier, the mastery tests are used to make a decision whether an examinee passes or masters a test or fails the test. Such decisions will be called classification decisions and the passing point will be referred to as the cut-score or cut-point in the following sections. Although there are several methods available for making classification decisions, for the present study, the only method used was sequential Baye's procedure (Owen, 1975). Since the method requires the estimated ability to be compared with the latent passing score to make a classification decision, the results are influenced by the item selection only and not by the method of scoring the test (Kalohn & Spray, 1998).

According to this method, probability of mastering a test (PM) was calculated and compared with the pass decision level. If this probability was greater than the decision level, 0.5 in this case, the examinee was classified as a master. In case of variable length test, the test terminated when either of the following criteria was satisfied:

(a) PM < lower limit of confidence region

(b) PM > upper limit of the confidence region

In this case the lower limit was employed as 0.1 so the test stopped when PM was greater than 0.9 or less than 0.1. If the examinee reached the maximum limit for the number of items but none of those criteria was satisfied, the classification method was similar to the fixed length test. In other words, the decision was based on the most recent update of the mastery probability (Kalohn & Spray, 1998).

The cut-score information for AICPA tests is provided for the overall tests. The raw scores are converted to a score scale of 1-100 and the passing score for each test is 75. The classification information is also available for each of the tests. In order to obtain an estimate

of the cut-score for the multiple-choice items, the classification information was used for each administration of the test. The derived cut-score information from each ability distribution is presented below:

Table 7: Derived Cut-Scores

	Audit			**ARE**		
	1996	1997	1998	1996	1997	1998
Derived cut-score	0.52	0.54	0.49	0.52	0.64	0.57

The average derived cut-scores were, therefore, taken as 0.52 and 0.58 for Audit and ARE respectively on the IRT ability scale.

The mastery tests, like performance (achievement) tests were also being simulated at two different test lengths. First, fixed length and then variable length tests were administered to the 1200 examinees. In case of variable length CAT, the examinees could take a minimum of 25 items and a maximum of 45 items. Those limits were chosen to depict half of the test length for paper and pencil version of the tests.

A significant aspect of these analyses was to look at the effect of rushed guessing on the classification decisions in a licensure examination. The effect of guessing on the ability estimation as well as the accuracy of classification decisions was analyzed for both fixed and variable lengths tests.

Analyses

In order to look at the differences between true and estimated ability estimates, Root Mean Square Error (RMSE) and Bias indices were computed. The RMSE index is defined as the standard deviation of estimated ability around true ability. Bias, on the other hand

provides us with a sense of direction of the estimated abilities relative to the truth. Following is the mathematical representation of the two indices:

$$RMSE = \sqrt{\frac{\sum_{j=1}^{N}(\hat{\theta}_i - \theta)^2}{N}}$$

$$BIAS = \frac{\sum_{j=1}^{N}(\hat{\theta}_i - \theta)}{N}$$

Here, N is the total number of examinees. The values for those indices are presented in both tabular and graphical formats for various levels of aberrance. For each administration of the test, average test information was computed at various ability levels for different response patterns. This provided us with an idea of how well the test was targeted to the examinee ability levels. For each aberrant condition, various plots were produced for the estimated ability against the items currently administered. This was repeated for several examinees at a variety of ability levels. Examinees with the same true abilities (which is frequent) were considered as the replications of the estimation at a particular ability level thus giving us a clear picture of the estimation process.

It would also be helpful to look at an index of pool utilization. Although a lenient exposure control method was used, it'd be helpful to look at the effect of aberrance on exposure rates and thus the pool utilization. We could hypothesize that greater aberrance could lead to increased Skewness in the exposure rates. One such index as proposed by Chang & Ying (1999) is a chi-square index of pool utilization that provides a measure of Skewness of exposure rates (Robin, 2000). The index is defined as, (to be included)

$$\chi^2 = \frac{\sum_{j=1}^{N}(er - L/N)}{N}$$

where *er* is the exposure rate, *L* is the number of items administered, *N* is the number of items in the pool. Although, the index cannot be relied upon in isolation, that is, without looking at exposure rates, it's used in this study to look at the general patterns of pool utilization for different guessing behaviors.

The information provided by the pool that was available for an examinee before the item selection began, was also examined.

Results

Results Based on Simulations

The first round of simulations was carried out to generate responses on computer adaptive tests for 1200 examinees at two test lengths of 30 and 75 items. An item pool of 600 items was simulated for these analyses. The examinees were made to guess after a certain percentage of items had been administered to look at the effect of guessing on the response patterns and the final ability estimates. The examinees were made to guess after 90% of items had been administered to simulate examinees that guess later in the exam. On the other hand, guessing after 25% of test administration depicts the response patterns for examinees that are extremely slow test takers or have very low ability and start guessing very early in the test. Figures (13) thru (17) in appendix B demonstrate a computer adaptive test administration for a low ability examinee when he/she started guessing at several points in time, ranging from very early to later in the test. Figures (18) to (27) show the same results for examinees with middle and higher levels of proficiency. For the purpose of these analyses, theta levels of –1.83 and 1.83 were arbitrarily chosen to depict lower and higher levels of ability, respectively. Theta level of 0.1 was chosen to simulate responses for

examinee with middle ability level. The following table displays the ability levels and the mid-point of the corresponding ability interval (the interval size was smaller around the mid-ability compared to the tails of the population to simulate a normal distribution of examinees):

Table 8: Ability Levels

Ability Levels	Mid-Point	Ability	Mid-Point
1	-1.83	7	0.10
2	-1.15	8	0.31
3	-0.80	9	0.54
4	-0.54	10	0.80
5	-0.31	11	1.15
6	-0.10	12	1.83

Figures (28) thru (42) in appendix B demonstrate the same analyses, however, the first set of analyses was performed for a test length of 30 while the second set for a test length of 75 items. The test length of 75 items reflects the test lengths for the various sub-tests for AICPA exam (the four sub-tests consist of 60 to 75 items). The plots indicate the estimated ability and item difficulty (vertical axes) after each item has been administered (horizontal axis). The vertical dashed line indicates the point after which an examinee starts guessing while the horizontal line is drawn across the true ability estimate of the examinee.

The results showed that the ability estimation considerably improved when the test length was increased from 30 to 75 items. An important finding in both cases was that the adverse effects of guessing were significant for middle and high ability examinees. Those effects were highly noticeable when the examinees started guessing at the earlier stages of the test. If we looked at the examinees that started guessing halfway through the test, the difference between the true and final estimated ability was a fraction of a point for a low ability examinee while one and two point difference was observed for middle and high ability

examinees respectively. Increasing the test length improved the estimation for the high ability examinee when he/she started guessing at a later point in the test. The test length didn't prove to have much of an effect on the estimation in general once the examinees started to guess. This result is demonstrated in figure (43) where root mean squared error in the ability estimation over 1200 examinees are plotted against the guessing points for the two test lengths. Increasing the test length actually resulted in higher RMSE and Bias indices.

The effect of guessing on the actual examinee responses is shown in figures (44) thru (48) in appendix B. The figures show an example of the way a pattern of responses could change when the examinee guesses. Here the true ability of the examinee was moderately high, hence a number of correct responses were obtained when the examinee did not guess under the "null" condition. The responses were adversely affected (more 0s than 1s) when the examinee rushed after a certain number of items had been administered.

One of the hypotheses that we could also formulate from our knowledge of the CAT is that if a large number of examinees are simultaneously taking a CAT from the same item pool, ability estimation for the group could be affected. Since the utilization of the pool is disturbed and distracted from the way it was supposed to be utilized, the quality of estimation could be affected in general. Various proportions of examinees that took the test about the same time were thus made to guess at various points in the test. Figure (49) in appendix B illustrates the fact when 60% of the examinees were flagged or were made to guess. For those flagged examinees, several scenarios were simulated as described below:

Scenario A: 20% of flagged examinees guess after 25% of items have been administered; 80% guess towards the end (after 90% of items have been administered)

Scenario B: 20% of flagged examinees guess after 50% of items have been administered; 80% guess towards the end

Scenario C: 20% of flagged examinees guess after 75% of items have been administered; 80% guess towards the end

Scenario D: All examinees guess towards the end of the test

For all the above scenarios, the root mean square errors in the final ability estimates were plotted for each situation as shown in figure (49). The results reiterate the fact of how rushed guessing starting at early stages in a CAT could influence the final ability estimates. A very large population of examinees who guess towards the end of the test could also influence the pool utilization such that the very few examinees that started guessing early might end up with very poor ability estimates. This hypothesis could be true if there was a lack of high quality easy items available for guessing examinees. The hypothesis will be examined more carefully in the next phase of study. Figures (50) and (51) in appendix B depict the values for average test information that the test would provide assuming that the examinees had rushed into guessing. Out of 1200 examinees that took the CAT about the same time, all of them guessed at one point or the other in figure (50). On the other hand, figure (51) shows that only 30% of the 1200 examinees guessed at those points. As expected the test information is significantly affected when a higher number of people guess, however the differences between the two situations are significant when guessing took place early in the test. The small percentage of population who guessed earlier had a serious impact on the average test information. The amount and pattern of guessing across the 12 ability levels stayed almost similar in both situations when the guessing took place later in the test.

Results for Proficiency Testing Using AICPA Parameters

The results for CAT simulated with AICPA parameters were based on the analyses performed for the simulated parameters. The root mean squared errors (RMSE) were used to examine at the estimation accuracy. Figures (52) to (61) in appendix C depict the distribution of examinees falling in a certain ability interval based upon true and estimated ability. A significant drop was observed in the number of examinees in the higher ability intervals as they started to guess. Same analyses were repeated for other types of tests (mastery etc). The figures for those tests are shown together so that readers can compare the figures when a reference is being made to those tests later in this chapter.

Figures (67) through (69) in appendix C show the Root Mean Squared Errors plotted against the various guessing behaviors for various ability levels at test lengths of 30 and 75 items. Figure (70) indicates the overall RMSE while the rest depict the plots for low, medium and high ability examinees. For a 75-item CAT, the error in estimation increased from 0.2 when there was no guessing to 0.3 when guessing was introduced towards the end, 0.7 when guessing began after 75% of the test was administered, 2.5 after 50% and 3.6 when guessing began very early. For a 30-item CAT, these values were 0.3 for no guessing situation, 0.4 when examinees guessed towards the end, 0.8, 2.0 and 3.3 for the respective guessing behaviors. Also shown in figure (70), these values were slightly higher for a 30-item test when guessing was introduced in the later part of the test while lower when guessing began earlier. When the RMSE errors were examined for examinees at various ability levels, it was found that the errors followed similar patterns for the two test lengths. An exception to this was the case of low ability examinees for whom the RMSE values constantly remained higher

for the 30-item test. The following table presents the values of RMSE for the three ability levels:

Table 9: Error in Estimation (RMSE) for Audit Sub-Test

	RMSE for a 75-item CAT				RMSE for a 30-item test			
Guessing	Overall	Low	Medium	High	Overall	Low	Medium	High
25%	3.61	1.45	3.67	5.40	3.28	1.69	3.14	4.70
50%	2.48	1.22	2.51	3.64	2.03	1.27	2.05	2.89
75%	0.66	0.49	0.68	0.94	0.75	0.74	0.77	0.88
90%	0.29	0.35	0.26	0.41	0.39	0.50	0.36	0.44
NG	0.19	0.30	0.18	0.20	0.27	0.36	0.26	0.24

The RMSE values were very similar at each ability level when the examinees did not rush into guessing. The errors were constantly higher for the high ability examinees and their differences from the middle and low ability examinees increased as examinees guessed early on. The examinees with middle ability levels were lower than the high ability examinees but higher than the low ability examinees in terms or error in estimation.

Bias in estimates is represented in table 10 (see figures 71 to 74 in appendix C).

Table 10: Bias in Estimates for Audit Sub-Test

	Bias for a 75-item CAT				Bias for a 30-item test			
Guessing	Overall	Low	Medium	High	Overall	Low	Medium	High
25%	1.83	1.13	1.89	2.31	1.74	1.23	1.72	2.13
50%	1.45	1.04	1.50	1.83	1.30	1.02	1.31	1.59
75%	0.75	0.59	0.75	0.95	0.75	0.70	0.74	0.88
90%	0.44	0.37	0.39	0.60	0.48	0.52	0.44	0.59
NG	0.09	0.22	0.04	0.09	0.13	0.26	0.09	0.17

The table shows that the bias increased significantly as soon as the guessing was introduced. Although, the RMSE values were negligibly small when examinees guessed towards the end, the bias in estimates was large. The largest amount of bias was observed in high ability examinees while the smallest amount of bias was observed for low ability examinees. For low ability examinees, bias was higher for the shorter test when guessing was introduced later

in the test. For medium and high ability examinees, difference in bias was negligible for the two test lengths when examinees guessed later. The differences, however, increased when examinees guessed earlier.

Figures (75) to (104) in appendix C show the administration of a proficiency CAT to a typical low, medium and high ability examinee for Audit at two test lengths. The significant drop in the estimates for high ability examinees once they guessed early, explains the high values for RMSE. The accuracy was somewhat lost when examinees guessed towards the end; the loss was greater for high ability examinees. Another significant finding was that the estimates decreased significantly for middle ability examinees once they guessed after 75% of the test had been administered. The estimates decreased further when guessing began after half of the test was administered. The estimates, however, remained more stable compared to those for low and high ability examinees.

Figures (105) and (106) represent the average test information at 12 ability levels for various guessing behaviors at two test lengths. As shown in figure (105), the test provided maximum amount of information for middle to high ability examinees and minimum amount of information at the tails of the distribution. Similar pattern was observed when a shorter test was administered, however, as expected, the information was much lower than the 75-item test. The information stayed much more stable across the ability levels when compared with longer test in both guessing and non-guessing scenarios. When examinees guessed later in the test, the information was lost at most of the ability levels except at higher ability levels. In case of ARE, a difference was that the information did not drop at the upper most tail of the distribution as was the case in Audit. When we look at figures (107) and (108) for the average pool information at various ability levels, similar patterns were observed.

An interesting aspect of the study was to look at the average information that the pool provided before item selection algorithm began for each examinee. As mentioned earlier, the aberrance in examinee behaviors might have an adverse effect on the pool configuration. The selection of unusual number of easy or difficult items during the time when examinees rush into guessing could result in a less informative pool for various ability levels. For this purpose, Fisher's information was computed for each item in the pool. Information for each item was then summed to obtain the total amount of pool information that was available for item selection at the beginning of a CAT.

In the previous section, reference was made to figures (107) and (108) for pool information. Looking at the same figures, it was found that guessing also affected the amount of information that pool provided for item selection. An interesting finding was that early guessing resulted in a pool that provided maximum amount of information at the uppermost end of the ability distribution. This finding was specifically apparent when the examinees were administered a shorter test. An explanation for this might lie in the fact that exposure rates are subject to change significantly when examinees guess very early in the test. If each examinee guesses early, easier items must get utilized very quickly. At this point, it is useful to look at the pool utilization index. Figures (109) thru (110) depict the plots for such index for two test lengths for several guessing scenarios. A slight increase was observed in the index when guessing increased. The value of the index was higher when examinees were administered a shorter test indicating more Skewness in the exposure rates for items in the pool.

Results for Proficiency Testing with Content Constraints using AICPA parameters

The results for proficiency testing remained very similar when content constraints were introduced in the test. The results indicated that the content constraints did not seem to have much effect on the aberrance. The existence of sufficient number of easy items in the pool for each content area simplified the complexity of item selection that could arise when guessing was introduced. The plots for average test information for various content areas in Audit are depicted in figures (113) to (116) in appendix C . Similar plots for a 30-item test are shown in figures (117) through (120). The overall average test information is presented in figures (121) and (122). The total test information for each examinee was re-scaled as the number of items in each content area was variable. The figures indicate that a large amount of information provided by the test was attributed to the first content strand.

In order to look at the effect of guessing when some content areas have fewer easy items than others, further simulations were conducted. The difficulty parameter for each item in the first content strand was increased by an arbitrarily chosen constant (1.2). The change in RMSE and Bias indices by changing the difficulty parameter for a single content strand was negligible, although that content area had the largest representation in the test. The plots for RMSE indices are shown in Figures (125) to (128) in appendix C.

The average test information for the various content strands is shown in figures (129) through (136) in appendix C. The average test information was significantly affected in the first content strand. A noticeable drop in the information was observed at all ability levels except for examinees with abilities in the highest range. An interesting finding was that the information for other content areas increased for a wider ability range when examinees guessed earlier, even though configuration of the pool was not altered for those content areas. The information, in this case, decreased for the higher most ability levels.

Results for Mastery Testing Using AICPA Parameters

The classification decisions were first examined for the fixed length tests for Audit and ARE at the test lengths of 30 and 75 items. Table 11 presents the results for Audit for the test length of 75 items indicating the total number of people passed based on the true ability and then based on the estimated ability. The table also shows the overall percentage of people who were classified correctly versus those classified incorrectly as well as the percentage of misclassifications. Each of these results was then examined for the various guessing scenarios to look at the effect of guessing on classification. Table 12 on the other hand shows the classification decisions broken down by ability level. The analyses indicated that out of 1200 examinees 400, examinees passed the Audit examination if the decisions were based upon their true abilities. The number of people who passed the test was reduced to 371 when the decisions were based on the estimated abilities. When the numbers were broken down by ability, it was

Table 11: Classification of Masters/Non-Masters (Fixed Length Audit--75 items)

Guessing points During CAT	People Passed True	People Passed Estimate	People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
No Guessing	400	371	1139	61	0.05
After 90%	400	277	1071	129	0.11
75%	400	149	949	251	0.21
50%	400	2	802	398	0.33
25%	400	0	800	400	0.33

observed that examinees only in the higher ability levels (>=0.54) were able to pass the test. Based on the estimated ability, 16 out of 371 people who were originally classified in the middle ability levels were able to pass the test. The largest reduction in the number of people who passed the test based on the estimated ability was observed at the cut-point. In other words, the largest number of misclassifications was observed around the cut-point.

Table 12: Percentage of Correctly Classified at each Ability Level (Fixed length Audit--75 items)

Ability	People Passed		Percentage of Correctly Classified				
Levels	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100
2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	0	100	100	100	100	100
7	0	2	**98**	100	100	100	100
8	0	14	**86**	**97**	100	100	100
9	100	58	**58**	**18**	**0**	**0**	**0**
10	100	97	**97**	**60**	**12**	**0**	**0**
11	100	100	100	**96**	**42**	**0**	**0**
12	100	100	100	100	**95**	**2**	**0**

The results were then analyzed for various guessing behaviors. As shown in table 12, it was found that the number of examinees who passed the test, dropped by approximately 25% when the examinees started guessing after 90% of the items had been administered. The number dropped by 60% when the examinees started guessing after 75% of the items had been administered and by 99.5% when the examinees started guessing very early in the test. In terms of the classification decisions, the number of people that were misclassified when there was no guessing, doubled when examinees started guessing after 90% of the items had been administered. The number of misclassifications increased by 4 times when guessing started after 75% of the test length and by approximately 6 to 7 times when examinees guess earlier. Out of the total population, the percentage of misclassifications increased from 5% to 11% for late guessing and 33% for earlier guessing.

As shown in table 12, the percentage of correctly classified increased from 98% to 100% for middle ability of 0.1 and from 86% to 97% for middle ability of 0.31 when examinees started guessing later in the test. However, the percentage significantly dropped

(58% to 18%) for examinees at or slightly above cut-score. In other words, the accuracy of classification increased for examinees slightly below the cut-score once they started to guess. When the examinees started to guess earlier, the percentage of correct classifications dropped to 0% for examinees at or above cut-score.

Table 13 depicts the results from the similar analyses for Audit when the test length was reduced to 30 items. The overall number of people who passed the test with lesser number of items decreased. However as the examinees started to guess randomly at a certain point in the

Table 13: Classification of Masters/Non-Masters (Fixed Length Audit--30 items)

Guessing points during CAT	People Passed		People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
	True	Estimate			
No Guessing	400	366	1116	84	0.07
After 90%	400	270	1054	146	0.12
75%	400	192	986	214	0.18
50%	400	23	823	377	0.31
25%	400	1	801	399	0.33

test, the number of examinees who passed the test increased when compared with the guessing behaviors in a longer CAT. For example, when examinees guessed after 75% of the 75-item test was administered, the number of people who passed the test was 149 compared to 192 people on a 30-item test. When guessing occurred after half way through the test, the number of examinees passing the test increased from 2 on a 75-item test to 23 on a 30-item test. The incorrect classifications were therefore less frequent in the case of a 30-item test when examinees did not guess or guessed towards the end.

Table 14: Percentage of Correctly Classified at each Ability Level (Fixed length Audit--30 items)

Ability Levels	People Passed		Percentage of Correctly Classified				
	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100

2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	1	**99**	100	100	100	100
7	0	6	**94**	**99**	**99**	100	100
8	0	18	**82**	**93**	**98**	100	100
9	100	53	**53**	**22**	**10**	**0**	**0**
10	100	89	**89**	**52**	**22**	**1**	**1**
11	100	99	**99**	**88**	**61**	**6**	**0**
12	100	100	100	100	**96**	**16**	**0**

However, the situation was reversed when guessing occurred earlier. When broken down by ability levels, it was found that the decrease in the degree of misclassification or in other words increase in the percentage of correctly classified was observed for ability levels slightly below the cut-point. The degree of correct classification was decreased at ability levels at or slightly above the cut-point. As in the case of 75-item test, the classification decisions were accurate for very low and very high abilities.

When the examinees started to guess towards the end of the test, the differences between the proportions of correctly classified for late-guessing and no-guessing scenarios remained stable for the two test lengths. The only exception was observed for very high ability examinees, where the proportion dropped by 11% in case of a 30-item test compared to 4% drop in the 75-item test. When guessing occurred at the beginning of the last quarter of the test (after 75%), the drop in the correct classification rates from no-guessing scenario was much larger for a longer test for high ability examinees. This was due to the fact that the number of correctly classified was higher for those examinees when they guessed earlier on the shorter test. For the examinees that were slightly above the cut-point, situation was similar to the 75-item test, that is, the percentage of correct classifications increased when examinees guessed later and became accurate when guessed earlier. The classification

decisions for all examinees at or above the cut-score were inaccurate in both cases when guessing occurred very early in the test.

Table 15 indicates results for similar analyses when performed for a variable length test for Audit. As mentioned earlier, a stopping rule pertaining to the level of confidence in pass/fail decisions was employed in this case. The results were much

Table 15: Classification of Masters/Non-Masters (Variable Length Audit)

Guessing points during CAT	People Passed		People Classified Correctly	People Class. Incorrectly	Percentage of Misclassifications
	True	Estimate			
No Guessing	400	363	1133	67	0.06
After 90%	400	174	974	226	0.19
75%	400	85	885	315	0.26
50%	400	4	804	396	0.33
25%	400	1	801	399	0.33

similar to the fixed length when guessing occurred very early (after 25%) in the test or did not occur at all. Guessing at very early stage in the CAT resulted in highly inaccurate classification decisions for examinees at or above the cut-point in all cases. The results in this case, however, were much different from the fixed length test when examinees guessed later in the test. In terms of the number of people who passed when there was no guessing involved, results were very similar for the fixed and variable length tests. When the examinees started to guess after 90% of items had been administered, the number of passing examinees dropped from 363 to 174, compared to a drop of 371 to 277 examinees on a 75-item test and a drop of 366 to 270 examinees on a 30-item test. The overall percentage of misclassifications was very similar for all testing modes for no-guessing situation (5% for75 items, 7% for 30 items and 6% for variable length). The proportion of misclassified examinees increased from 11% on 75-item test and 12% on 30-item test to 19% on variable

length CAT when examinees started guessing towards the end. Similarly, this proportion was increased from 18% on 75-item test and 21% on 30-item test to 26% on variable length test.

When broken down by ability, it was observed that the accuracy of decisions suffered much more than the fixed length tests for examinees at or above the cut-score except those with the highest level of ability. The percentage of misclassification at cut-point decreased from 56% when the examinees did not guess, to 1% when guessing started after 90% of items had been administered. Similarly, these proportions decreased from 92% to 18% and 100% to 56% for the next higher levels of ability above the cut- score. The results for these two ability

Table 16: Percentage of Correctly Classified at each Ability Level (Variable Length Audit)

Ability	People Passed		Percentage of Correctly Classified				
Levels	True	Estimate	No Guess	After 90%	After 75%	After 50%	After 25%
1	0	0	100	100	100	100	100
2	0	0	100	100	100	100	100
3	0	0	100	100	100	100	100
4	0	0	100	100	100	100	100
5	0	0	100	100	100	100	100
6	0	0	100	100	100	100	100
7	0	3	**97**	100	100	100	100
8	0	12	**88**	100	100	100	100
9	100	56	**56**	**1**	**0**	**0**	**0**
10	100	92	**92**	**18**	**3**	**0**	**0**
11	100	100	100	**56**	**21**	**3**	**1**
12	100	100	100	**99**	**61**	**1**	**0**

levels were rather drastic when compared to fixed length CATs. The drop in the proportion of correctly classified people in the above mentioned guessing scenario was approximately double for the examinees with ability level slightly higher the cut-point. The drop in accuracy was approximately five times the drop for a 30-item fixed length test and approximately thirteen times for a 75-item test for examinees with abilities much higher than the cut-score.

It would be useful to look at the number of items that were attempted by the examinees before the test terminated. The following table shows the frequency of examinees that

Table 17: Number of Items Taken by Examinees at Various Ability Levels (Variable Length Audit Sub-Test)

	No Guess					Guessing Introduced after 90% of items				
	25	**26-30**	**31-35**	**36-40**	**41-45**	**25**	**26-30**	**31-35**	**36-40**	**41-45**
1	100	0	0	0	0	100	0	0	0	0
2	100	0	0	0	0	100	0	0	0	0
3	100	0	0	0	0	99	1	0	0	0
4	100	0	0	0	0	100	0	0	0	0
5	98	2	0	0	0	96	3	1	0	0
6	86	4	5	2	3	93	5	2	0	0
7	64	7	5	6	18	84	11	5	0	0
8	38	11	7	4	40	54	36	9	0	1
9	12	13	5	3	67	26	43	21	6	4
10	41	12	10	4	33	19	38	33	10	0
11	85	10	3	0	2	56	10	27	5	2
12	100	0	0	0	0	99	0	1	0	0
Total	**924**	**59**	**35**	**19**	**163**	**926**	**147**	**99**	**21**	**7**

	Guessing Introduced after 75% of items					Guessing Introduced after 50% of items				
	25	**26-30**	**31-35**	**36-40**	**41-45**	**25**	**26-30**	**31-35**	**36-40**	**41-45**
1	100	0	0	0	0	100	0	0	0	0
2	100	0	0	0	0	100	0	0	0	0
3	100	0	0	0	0	100	0	0	0	0
4	100	0	0	0	0	100	0	0	0	0
5	100	0	0	0	0	99	1	0	0	0
6	100	0	0	0	0	99	1	0	0	0
7	93	7	0	0	0	98	2	0	0	0
8	85	12	3	0	0	96	4	0	0	0
9	68	17	15	0	0	94	6	0	0	0
10	43	36	16	4	1	91	6	2	1	0
11	32	37	15	13	3	76	20	4	0	0
12	58	19	17	5	1	54	31	12	3	0
Total	**979**	**128**	**66**	**22**	**5**	**1107**	**71**	**18**	**4**	**0**

Table 17: (cont.)

	Guessing Introduced after 25% of items				
	25	**26-30**	**31-35**	**36-40**	**41-45**
1	100	0	0	0	0
2	100	0	0	0	0
3	100	0	0	0	0
4	99	1	0	0	0

5	100	0	0	0	0
6	100	0	0	0	0
7	100	0	0	0	0
8	99	1	0	0	0
9	100	0	0	0	0
10	99	0	1	0	0
11	98	2	0	0	0
12	97	2	1	0	0
Total	**1192**	**6**	**2**	**0**	**0**

attempted a certain number of items. The numbers of items are grouped into five classes of 25, 26-30, 31-35, 36-40, 41-45. The results indicated that examinees around the cut-point attempted lesser items once they started guessing. It was also observed that the earlier they guessed, the lesser items they attempted. However, looking at table 17, the classification accuracy was adversely affected for slightly higher ability examinees above the cut-point and not for the ones below that threshold. In the first block of the table, a significant observation was that 67% of the examinees closest to the cut-point attempted 41-45 items. This proportion dropped to 4% when guessing was introduced towards the end of the test. A large number of those examinees attempted 26-30 items.

The next phase of the analyses were performed to look at the estimation accuracy of the fixed and variable length mastery tests. As expected, the estimation accuracy remained very similar to the proficiency testing for both fixed and variable length tests. The results of mastery testing are presented in figures (137) to (150). An interesting fact was observed when we looked at the average information for a variable length test. The information monotonically increased till it peaked for examinees around the cut-score. After that point it became increasingly less till the uppermost ability level. This indicates that the examinees with abilities around the cut-score were presented most informative items. The information, in general, was decreased when compared with the fixed length mastery tests, being closer to

the information provided by 30-item test. Interestingly, the peak disappeared when the examinees started guessing. The information curve remained relatively flat over the ability levels when examinees guessed. As mentioned above, a large number of examinees took longer test when guessing was not introduced, while that number significantly decreased when guessing was introduced. The average pool information, however, followed a pattern very similar to a 30-item fixed length test.

Conclusion

The study shed light on some of the most important issues in computerized adaptive testing. The purpose of any assessment instrument is defied if the information obtained on that instrument leads to an incorrect decision. In adaptive testing, an incorrect decision at any point in the test can lead to serious discrepancies towards the end. Since each item or question that gets administered to an examinee has impact on the properties of the remainder of the test, any disruption in the test administration process is consequential.

The act of an examinee rushing into random guessing at any point in the test could result in misleading estimates of that examinee's proficiency. As serious as it is in proficiency or achievement tests, the problem of inaccurate estimation could be worse in Mastery testing. The declaration of examinees as Masters or Non-Masters on the basis of incorrect measures of their ability is no-doubt harmful.

The results of the study clearly indicate that the error in estimation increases significantly once the examinee rushes to finish the test. One could be misled into assuming that the low level examinees would be affected most by disruption in the item selection algorithm. The results of the study showed that the high ability examinees suffered most once

they ran out of time. In all cases, the error in estimates was lowest for low ability examinees, higher for middle ability examinees and highest for high ability examinees.

When examinees rushed later in the test (after 75%), a slight drop was observed in the estimates for high ability examinees. The estimates for low and middle ability examinees generally remained stable when they guessed towards the very end. For some low and middle ability examinees, the estimates even improved when they guessed. The inaccuracy became evident when examinees started guessing after 75% of the test had been administered. The inaccuracy increased as the point in time at which guessing was introduced moved earlier.

Increasing test length improved estimates by negligibly small amounts when guessing occurred later in the test. Increasing test length proved to have adverse effects on the estimation accuracy when the guessing started after almost 75-80% of the test had been administered. The negativity of increasing the test length was more significant for high ability examinees.

When the CAT was administered with content constraints, the error in estimation and the test information was not affected as long as enough items of varying difficulties existed in the pool. Same patterns of errors as well as test information were observed as in the case when the test didn't have content constraints. The average information was greatly affected when the pool lacked easy items in a content area and the examinees rushed to complete the test. In order to satisfy the content requirements, the algorithm was forced to select items for that content area but early guessing caused loss in information at all ability levels due to the lack of easy informative items. The situation got worse for very early guessers when the test length was increased as they

In case of mastery testing, the results were consequential. As guessing was introduced, people were incorrectly classified. The number of people who passed the test significantly decreased when guessing was introduced. The number of misclassifications was larger for the variable length test compared to the fixed length test. The results indicated that examinees around the cut-point suffered most when guessing was introduced for both fixed and variable length tests. In case of fixed length CAT, the number of incorrectly classified examinees was larger for longer tests.

When the results from simulations using simulated item parameters were compared to those from simulations using AICPA parameters, it was found that the estimation for early guessers was better in the former case.

Limitations

This study was limited in terms of several factors. The most significant limitation of the research was the imposition of time limits in the absence of a time-recording option. The timing conditions can be simulated more accurately if the examinees' response times are actually recorded by a built-in timer or a clock. Another limitation was the simplicity of the test design in terms of test content, item formats and item types. The study focused on multiple-choice items only and represented major content strands on the test.

Recommendations

The adverse effects of the interaction of test and examinee characteristics with response aberrance can be reduced to some extent in several ways. Based on the results of the study, several suggestions can be made to address the issue. One suggestion is the use of an index where response time adjustment is included in the item selection process. The time spent on an item could then be a part of the selection algorithm. Another possibility could be

to build the aberrance flags into the weighted deviations model. In other words, aberrant conditions could be controlled as part of the selection model.

In case of variable length mastery tests, increasing the minimum test length could prevent shorter tests to be administered to examinees. A minimum test length of 25 items proved to be problematic and resulted in numerous false classification decisions when aberrance occurred.

The use of Bayesian estimation with stronger priors has proven to provide better estimates than Maximum Likelihood estimation. The same finding was reinforced by this research. The estimates were largely affected by aberrance when the rushed guessing was introduced early in the test. In real testing environment, rushed guessing is observed towards the later part of the test for most of examinees depicting aberrant response patterns. Hence Bayesian estimation can prove to be a better way to estimate ability for majority of the population. It is also expected that for early guessers, the MLE would lead to estimates much further from the truth compared to Bayesian estimates. Further research is however needed to shed light on this result.

It would be imperative to conclude by emphasizing on the well-stated fact that creating richer pools can always reduce the gravity of the problem. Frequently occurring aberrance could lead to unexpected utilization of the pool, hence it is highly desirable to have a large pool with informative items at all ability levels.

References

American Council on Education (1995). Guidelines for computerized adaptive test development and Use in Education. Washington DC: Author.

Ackerman, T. (1987, April). The use of unidimensional item parameter estimates of multidimensional items in adaptive testing. Paper presented at the annual meeting of the American Educational Research Association Conference, Washington DC. (ERIC Document Reproduction Service No. Ed 284 901)

Bergstrom, B. & Garshon, R. (1992, April). Comparison of item targeting strategies for pass/fail computer adaptive tests. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA. (ERIC Document Reproduction Service No. Ed 400 287)

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Ed.), Statistical theories of mental scores (chapters 17-20). Reading, Massachusetts: Addison-Wesley.

Bracey, G. & Rudner, L. M. (1992). Person-Fit Statistics: High potential and many unanswered questions. Research Report 92-5. Office of Educational Research and Improvement (ED), Washington, DC.

Bradlow, E. T., Weiss, R. E. & Cho, M. (1998). Bayesian identification of outliers in computerized adaptive tests. Journal of the American Statistical Association, 94(443), 910-919.

Carlson, R. (1994). Computer adaptive testing: a shift in the evaluation paradigm. Educational Technology Systems, 22(3), 213-224.

Chang, H. (1996, April). A model for score maximization within a computerized adaptive testing environment. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Chang, H., Qian, J., & Ying, Z. (2000, April). *a*-Stratified multistage CAT with b-blocking. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Chang, H. & Ying, Z. (1997, March). A global information approach to computerized adaptive testing. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL.

Chang, H. & Ying, Z. (1999). *a*-Stratified multistage computerized testing. Applied Psychological Measurement, 23(3), 211-222.

Clark, C. (1976). Proceedings of the first conference on computerized adaptive testing. Washington DC: US Government Printing Office.

Dodd, B. G. (1990). The effect of item selection procedure and step size on computerized adaptive attitude measurement using the rating scale model. Applied Psychological Measurement, 14(4), 355-366.

Drasgow, F. & Parsons, C. (1983). Application of unidimensional item response theory to multidimensional data. Applied Psychological Measurement, 7, 218-232.

Eggen, T. H. (1999). Item selection in adaptive testing with the sequential probability ratio test. Applied Psychological Measurement, 23(3), 195-210.

Eignor, D. R., Stocking, M. L., Way, W. D. & Steffen, M. (1993). Case studies in computer adaptive design through simulation. Research Report RR-93-56. Educational Testing Service, Princeton, NJ.

Glas, C. A., Meijer, R. R. & Van Krimpen-Stoop, E. (1998). Statistical Tests for Person Misfit in Computerized Adaptive Testing. Research Report 98-01. Educational Science and Technology, University of Twente, Netherlands.

Green, B., Bock, R., Humphreys, L. & Reckase, M. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21(4), 347-360.

Guttman, L. A. (1944). A basis for scaling qualitative data, as described in Weiss, D. J. (1983).

Hambleton, R. & Swaminathan, H. (1985). Item response theory: principles and applications. Massachusetts: Kluwer Academic Publishers.

Hambleton, R., Swaminathan, H. & Rogers, J. (1991). Fundamentals of item response Theory. Newbury: Sage Publications.

Hambleton, R. (1983). Applications of Item Response Theory. Vancouver: Educational Institute of British Colombia.

Hambleton, R. K. & Pieters, P. M.. & Zaal, N. J. (1991). Computerized adaptive testing: theory, applications, and standards. In Hambleton, R. K. & Zaal, N. J. (Ed.). Advances

in educational and psychological testing: theory and applications (pp. 341-366). Massachusetts: Kluwer Academic Publishers.

Hick, W. E. (1951). Information theory and intelligence tests, as described in Kreitzberg et al. (1978).

Jensema, C. J. (1974). The Validity of Bayesian Tailored Testing. Educational & Psychological Measurement, 34(4), 757-66.

Kalohn, J. & Spray, J. (1998, April). Effect of item selection on item exposure rates within a computerized classification test. Paper presented at the annual meeting of the National Council on Measurement in Education. San Diego, CA.

Koch, W. & Reckase, M. (1979, April). Problems in application of latent trait models to tailored testing. Paper presented at the annual meeting of the National Council on Measurement in Education. San Francisco, CA. (ERIC Document Reproduction Service No. Ed 177 196)

Kogut, Jan. (1986). A review of IRT-Based Indices for Detecting and Diagnosing Aberrant Response Patterns. Report 86-4. Department of Education, Twente University, Netherlands.

Kogut, Jan. (1987). Detecting aberrant response patterns in the Rasch Model. Research Report 87-3. Department of Education, Twente University, Netherlands.

Kogut, Jan. (1987). Reduction of bias in Rasch estimates due to Aberrant Patterns. Research Report 87-5. Department of Education, Twente University, Netherlands.

Kreitzberg, C., Stocking, M. & Swanson, L. (1978). Computerized adaptive testing: principles and directions. Computers & Education, 2, 319-329.

Laurier, M. (1990, April). What can we do with computerized adaptive testing and what we cannot do?. Paper presented at the annual meeting of the Regional Language Center Seminar, Singapore. (ERIC Document Reproduction Service No. Ed 322 729)

Lawley, D. N. (1943). On problems connected with item selection and test construction, as described in Weiss, D. J. (1983).

Leucht, R. M., Nungester, R. J. & Hadadi, A. (1996, April). Heuristics based CAT: Balancing item information, content and exposure. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Lord, F. M. (1952). A theory of test scores. Psychometric monograph. No. 7.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental scores. Massachusetts: Addison-Wesley.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. New Jersey: Hillsdale.

McBride, J. & Wiess, D. (1983). Bias and Information of Bayesian Adaptive Testing (Research Report 83-2). Minnesota: Minnesota University.

McLeod L. D. & Lewis, C. (1996, April). Person-Fit indices and their role in the CAT environment. Paper presented at the annual meeting of the National Council on Measurement in Education. New York, NY.

Meijer, R. R. (1994). The influence of the presence of deviant score patterns on the power of a person-fit statistic. Research Report 94-1. Educational Science and Technology, University of Twente, Netherlands.

Meijer, R. R. & Sijtsma, K. (1994). Detection of aberrant score patterns: a review of recent developments. Research Report 94-8. Educational Science and Technology, University of Twente, Netherlands.

Mills, C. & Stocking, M. (1996). Practical issues in large scale high stakes computerized adaptive testing. Applied Measurement in Education, 9(4), 287-304.

Mosier, C. I. (1941). Psychophysics and mental test scores: the constant process, as described in Weiss, D. J. (1983).

Nering**,** M. L. (1995). The Distribution of Person-fit Using True and Estimated Person Parameters. Applied Psychological Measurement, 19(2) 121-29.

Parshall, C. G., Kromrey, J. D. & Hogarty, K. Y. (2000, April). Sufficient simplicity or comprehensice complexity? A comparison of probabilistic and stratification methods of exposure control. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Reise, S. P., Due, A. M. (1991). The influence of test characteristics on the detection of aberrant response patterns. Applied Psychological Measurement. 15(3), 217-226.

Revuelta J. & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. Journal of Educational Measurement, 35(4) 311-327.

Robin, F. (2000). Computer Based Testing Software (CBTS). Laboratory of psychometric and evaluative research. Amherst: University of Massachusetts.

Schartz, M. (1986, April). Measuring up in an individualized way with CAT-ASVAB: considerations in the development of adaptive testing pools. Paper presented at the annual meeting of the American Educational Research Association Conference. San Francisco, CA. (ERIC Document Reproduction Service No. Ed 269 463)

Stocking, M. (1987). Two simulated feasibility studies in computerized adaptive testing. Applied Psychology: An international review, 36, 263-277.

Stocking, M. (1996). An alternative method for scoring adaptive tests. Journal of Behavioral statistics , 21(4), 365-389.

Stocking, M. & Swanson, L. (1993). A Method for Severely Constrained Item Selection in Adaptive Testing. Applied Psychological Measurement, 17(3) 277-292.

Sympson, J. B. & Hetter, R. D. (1985). Controlling item exposure rates in computerized adaptive testing, as described in Stocking, M. & Lewis, C. (1996).

Thissen, D. (1990). Reliability and measurement precision in computerized testing. In Wainer, H. (Ed.). Computerized adaptive testing: A primer (pp. 161-185). New Jersey: Lawrence Erlbaum. Associates.

Traub, R. E. (1983). A priori considerations in choosing an item response model. In Hambleton, R. (Ed.). Applications of item Response theory (pp. 57-70). Vancouver: Educational Institute of British Colombia.

Traub, R. & Wolfe, R. (1981). Latent trait theories and the assessment of educational achievement. In Berhner, D. (Ed.). Review of Research in Education, 9, 377-435.

Wainer, H. (1983). Are we correcting for guessing in the wrong direction?. In Wiess, D. (1983). New horizons in testing (pp. 63-70). New York: Academic Press, Inc.

Wainer, H. (1990). Computerized adaptive testing: A primer. New Jersey: Lawrence Erlbaum Associates.

Wang, W., Wilson, M., & Adams. (1995). Item response modeling for multidimensional between-items and multidimensional within-items. Paper presented at the International Objective Measurement Conference. Berkeley, CA.

Weiss, D. (1982). Improving measurement quality and efficiency with adaptive testing. Applied psychological Measurement, 6, 473-492.

Wiess, D. (1983). New horizons in testing. New York: Academic Press, Inc.

Wiess, D. & Kingsbury, G. (1983). A comparison between IRT-based adaptive mastery testing and sequential mastery testing. In D. Wiess (Ed.). New horizons in testing (pp. 257-283). New York: Academic Press, Inc.

Wise, S. (March, 1997). Examinees issues in CAT. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. (ERIC Document Reproduction Service No. Ed 408 329)

van der Linden, W. J. (1986). The changing conception of measurement in education and psychology. Applied Educational Measurement, 10(4), 325-332.

van der Linden, W. J. & Scrams, D. J. & Schnipke, D. L. (1999). Using response time constraints to control for differential speededness in computerized adaptive testing. AppliedPsychological Measurement, 23(3), 195-210.

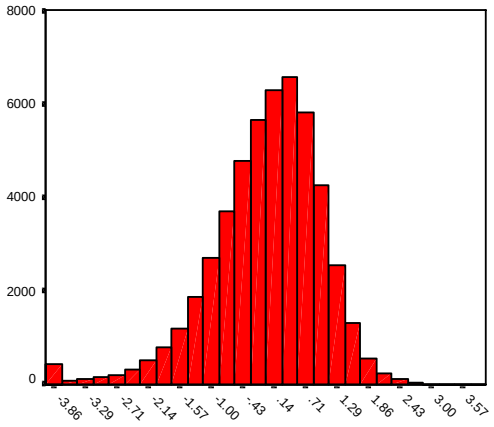
Yi, Q. & Nering, M. L. (1999). Simulating nonmodel-fitting responses in a CAT environment. Paper presented at the annual meeting of the National Council on Measurement in Education. Chicago, IL. (ERIC Document Reproduction Service No. Ed 427 042)

Yoes, M. E. & Ho, K. T. (April, 1991). The degree of Person Misfit on a nationally standardized achievement test. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

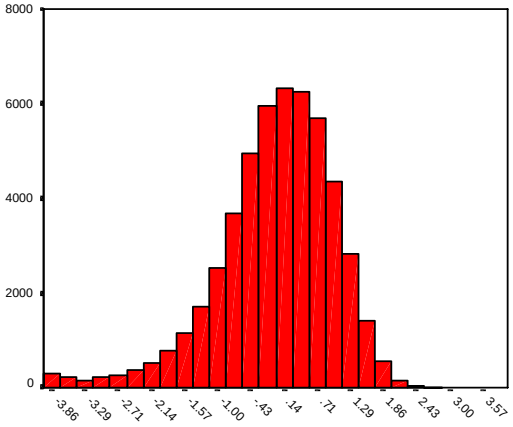
APPENDIX A

Ability Distributions for November Administrations of Audit and ARE

(1) 1996 Audit

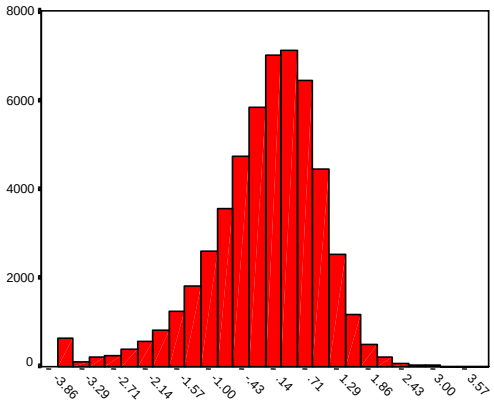


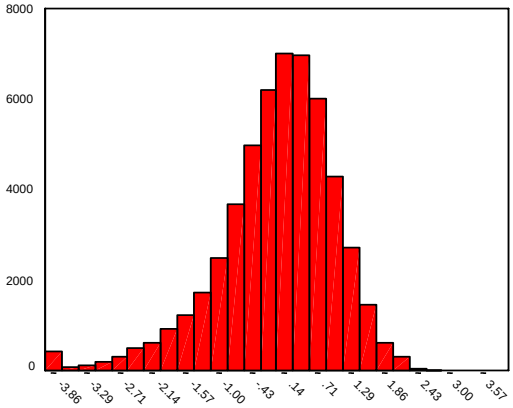
(2) 1996 ARE



(3) 1997 Audit

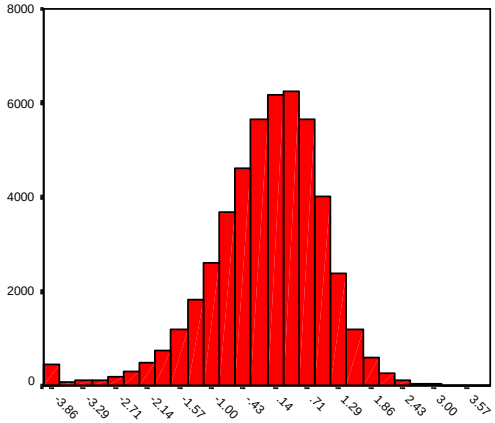
(4) 1997 ARE

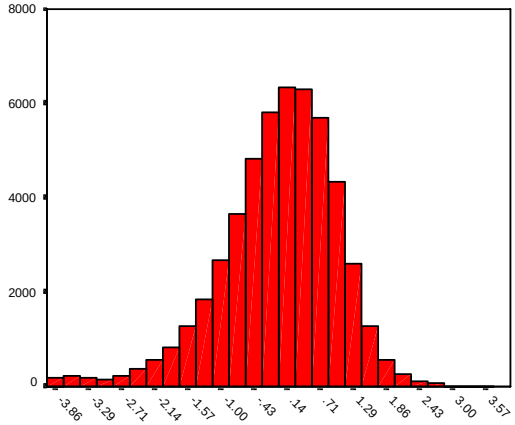




(5) 1998 Audit

(6) 1998 ARE

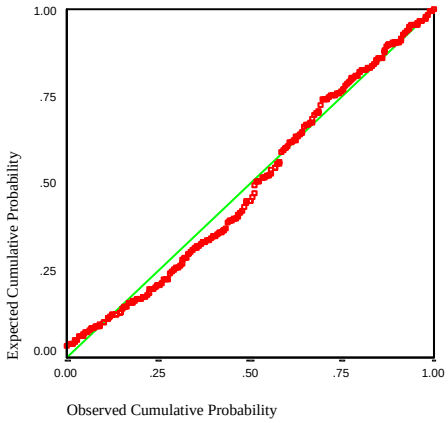


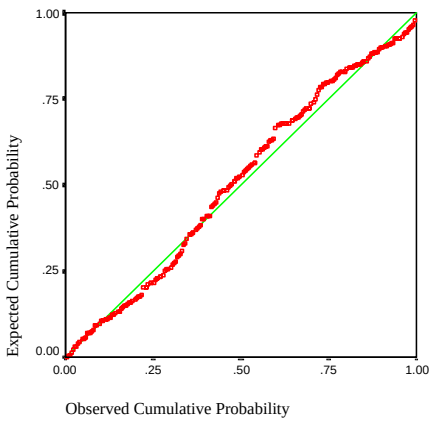


P-P Plots for AICPA Item Parameters

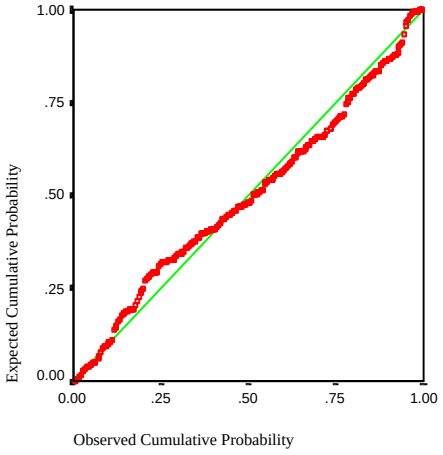
(7) Normal P-P for Audit a

(8) Log-Normal P-P for ARE a

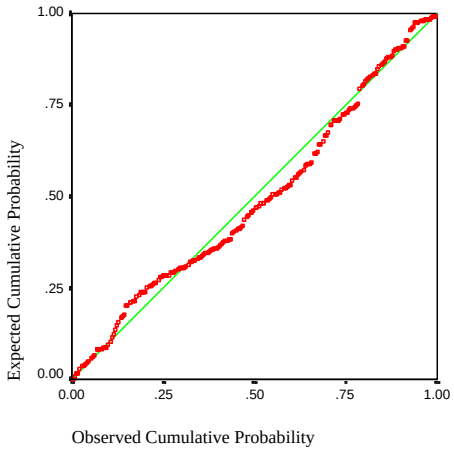




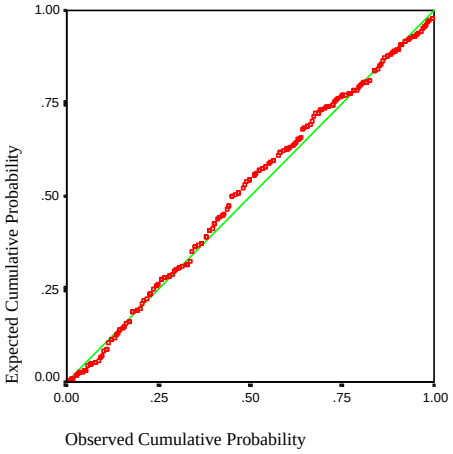
(9) Normal P-P for Audit b



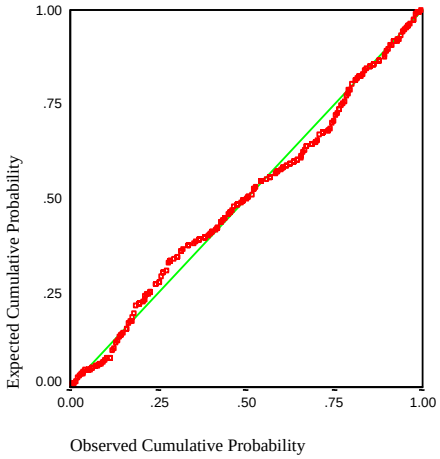
(10) Normal P-P for ARE b



(11) Normal P-P for Audit c



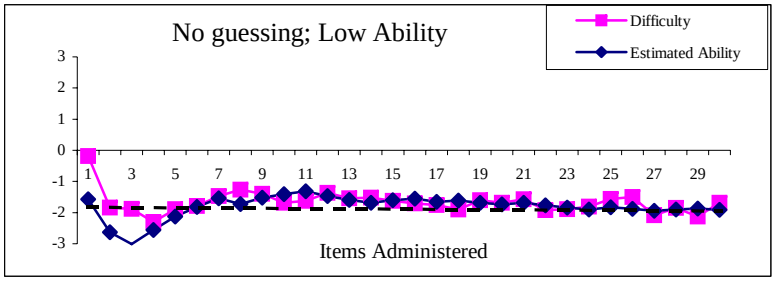
(12) Log-Normal P-P for ARE c



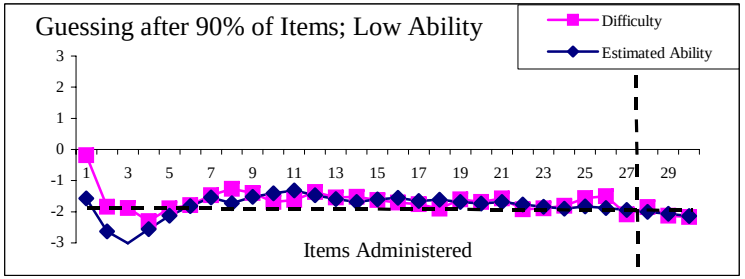
APPENDIX B

CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (30 items)

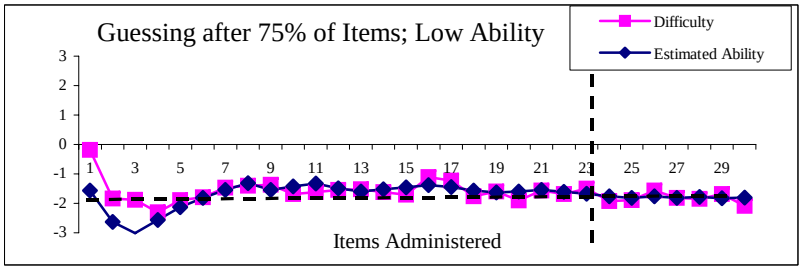
(13)



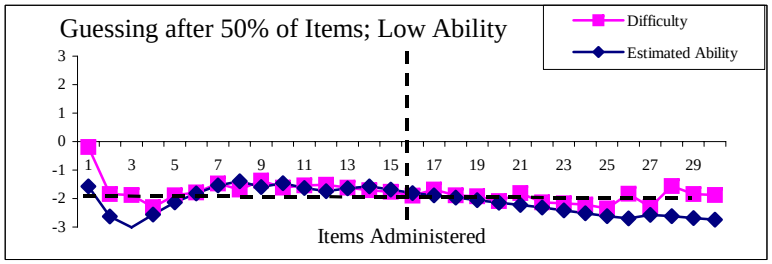
(14)



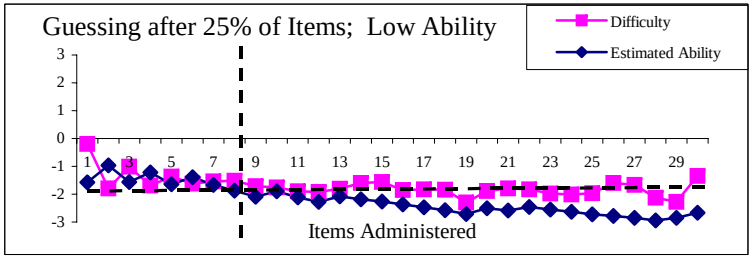
(15)



(16)

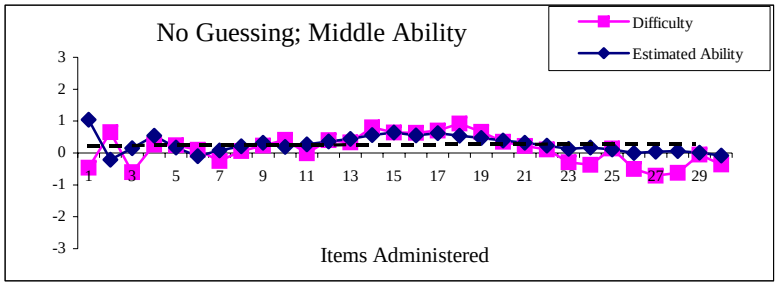


(17)

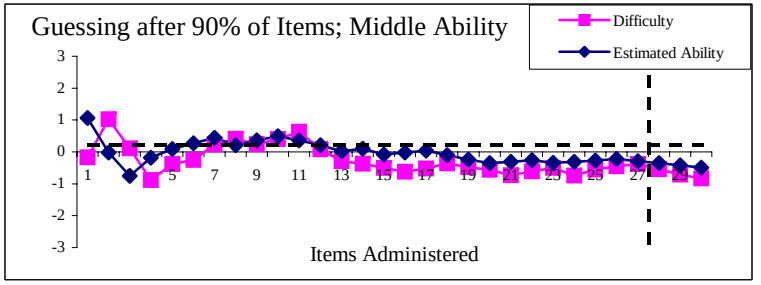


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (30 items)

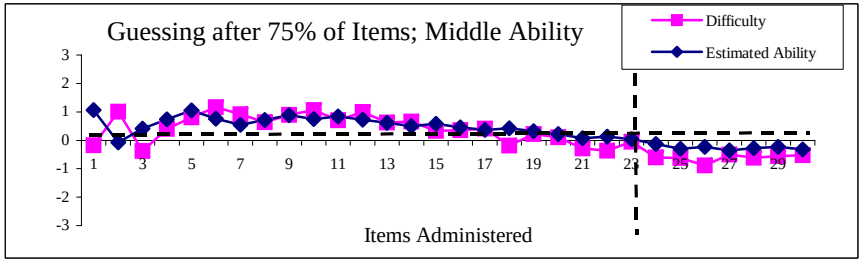
(18)



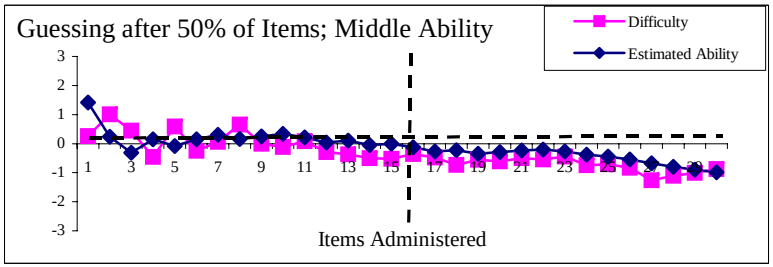
(19)



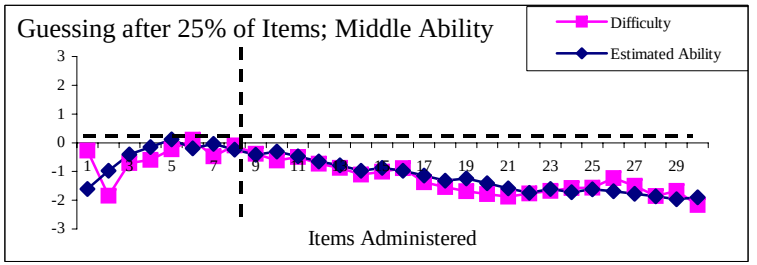
(20)



(21)

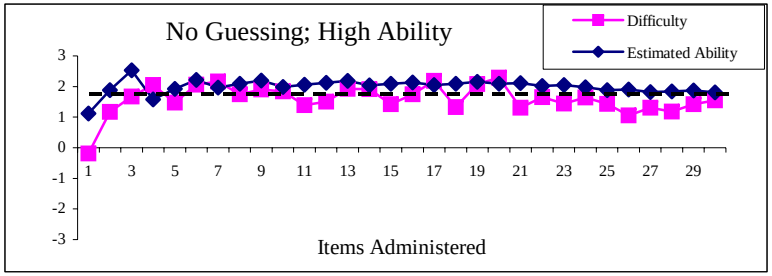


(22)

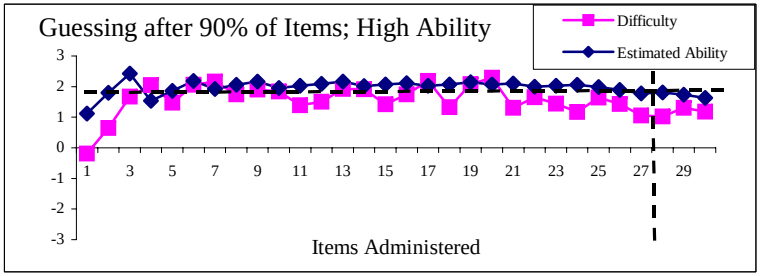


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (30 items)

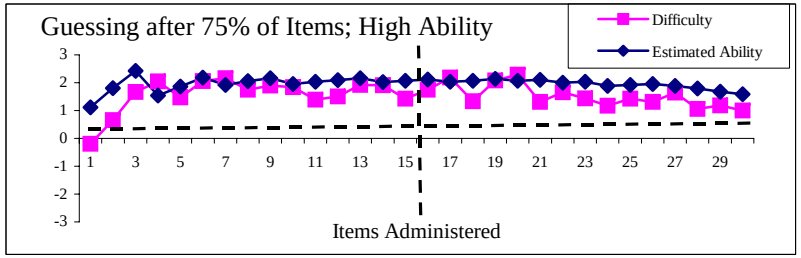
(23)



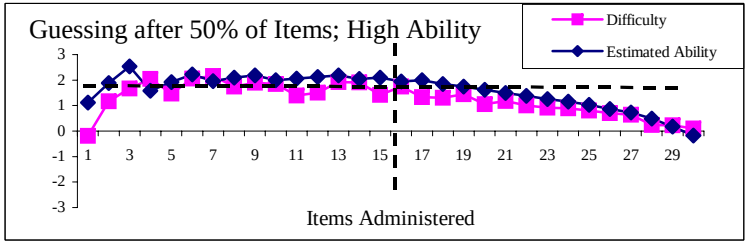
(24)



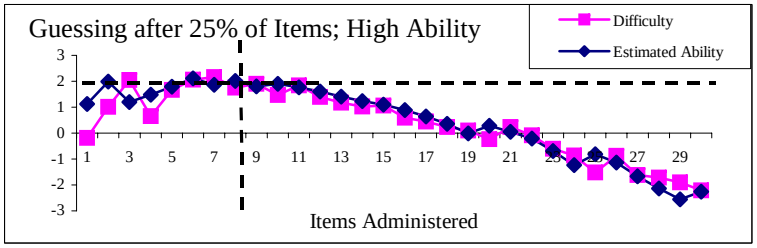
(25)



(26)

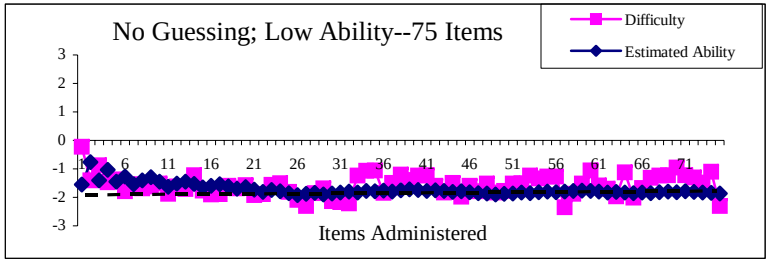


(27)

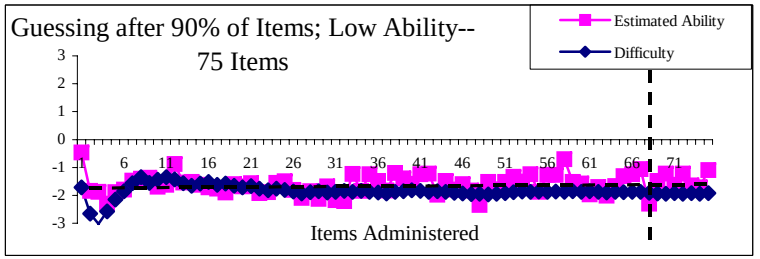


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (75 items)

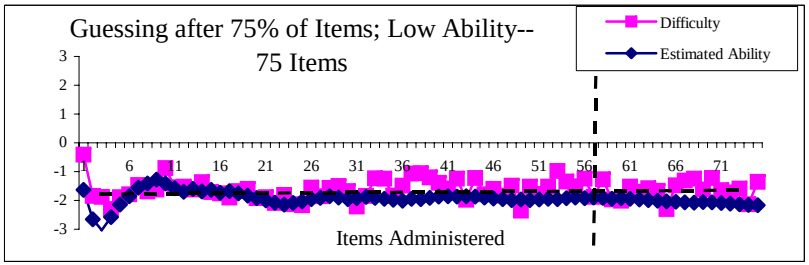
(28)



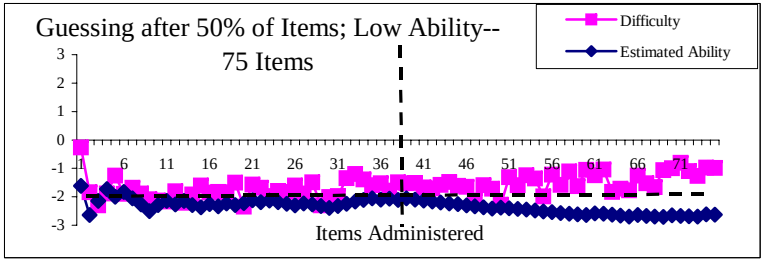
(29)



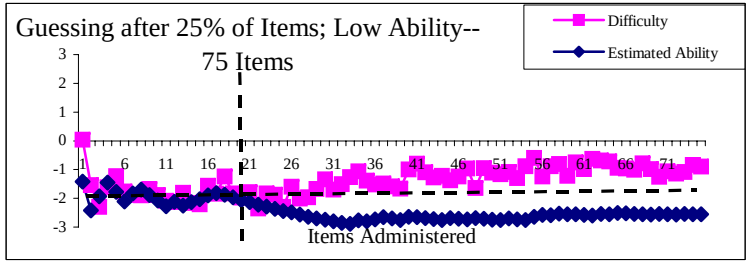
(30)



(31)

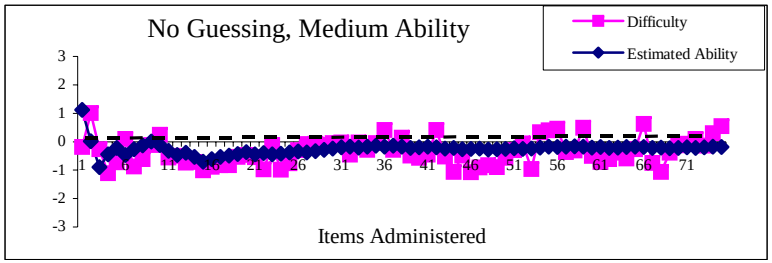


(32)

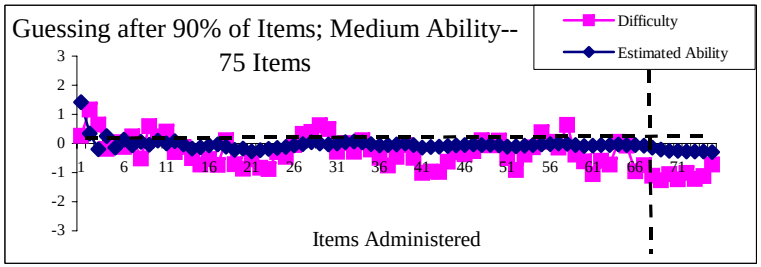


CAT Administration for a Medium Ability Examinee who Guesses at a Certain Point in the Test (75 items)

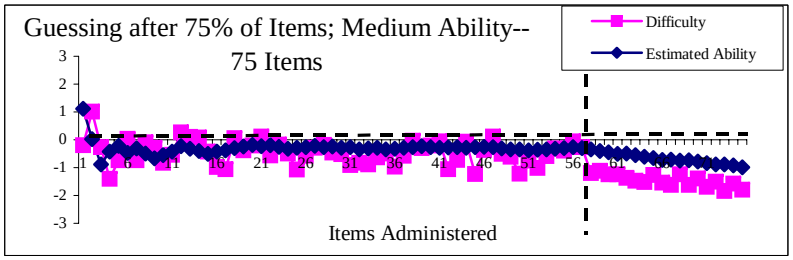
(33)



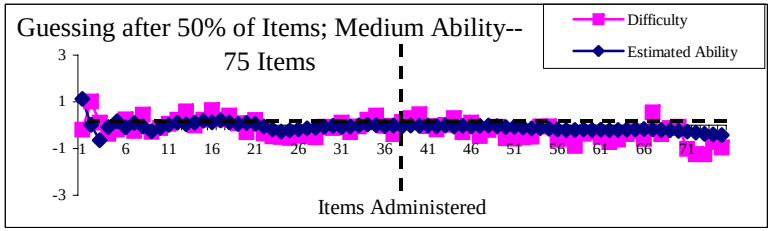
(34)



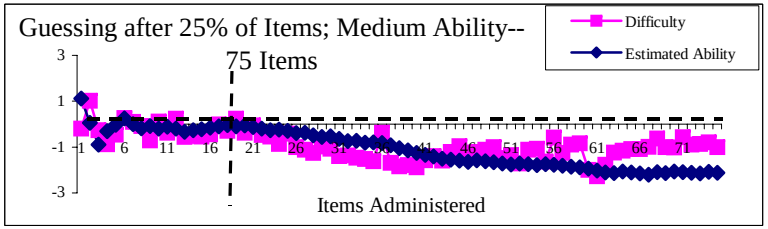
(35)



(36)

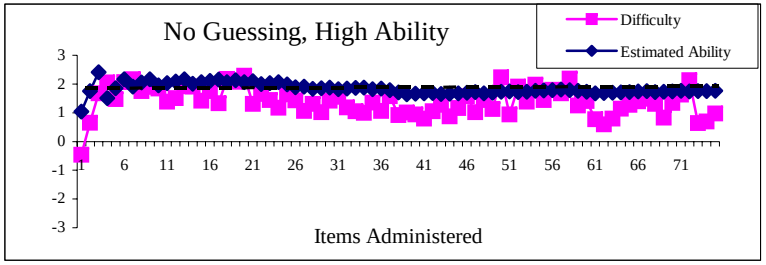


(37)

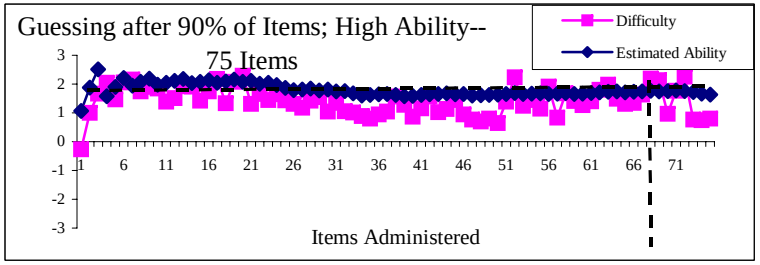


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (75 items)

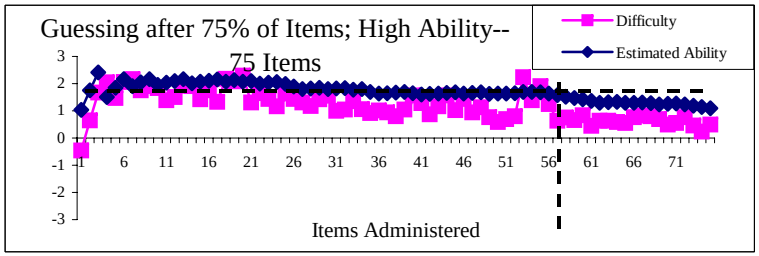
(38)



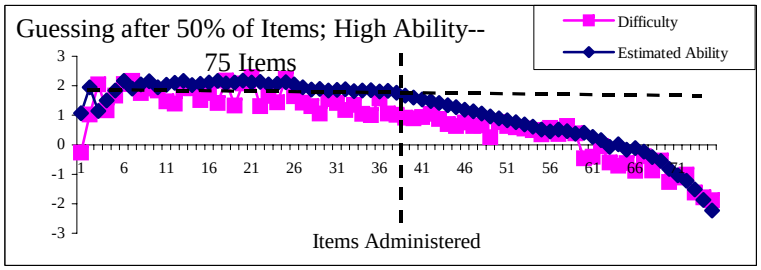
(39)



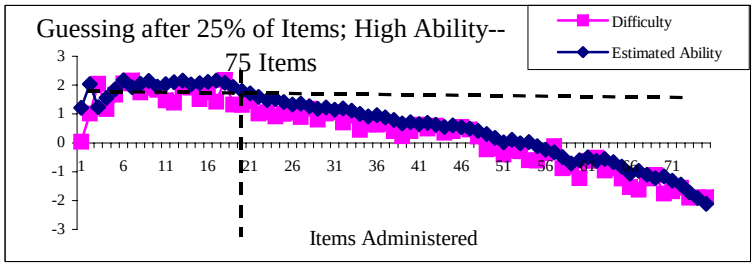
(40)



(41)

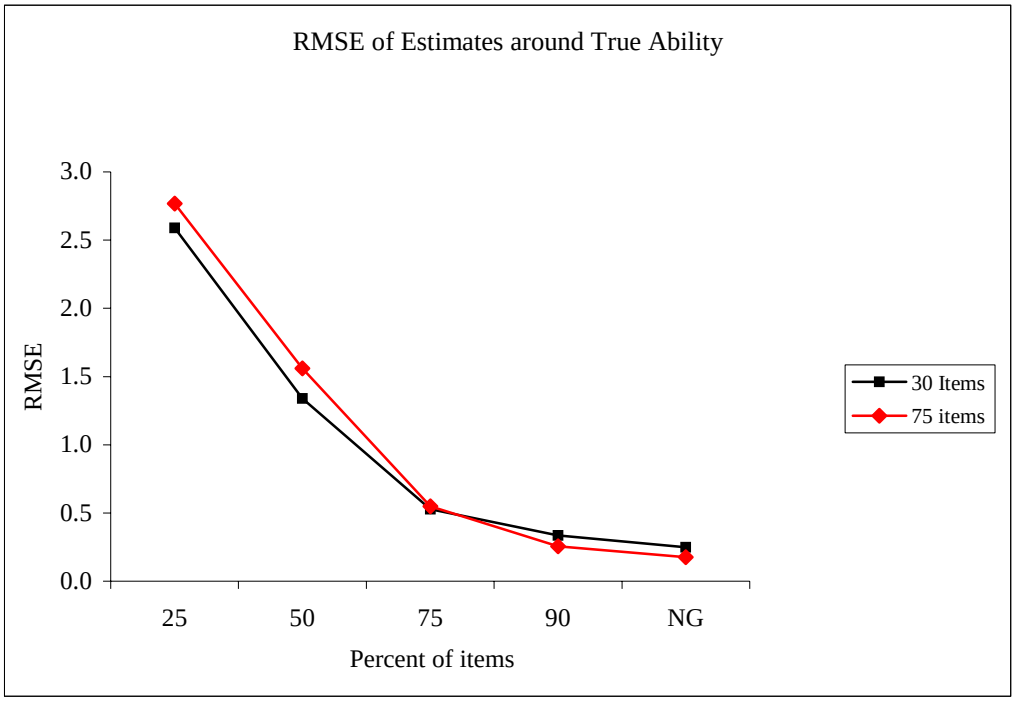


(42)



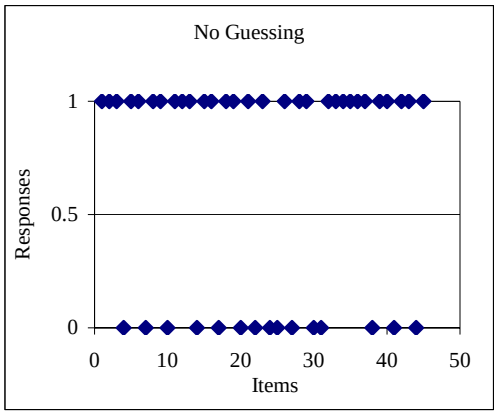
Error in Ability Estimation for Various Guessing Behaviors at Two Test Lengths

(43)

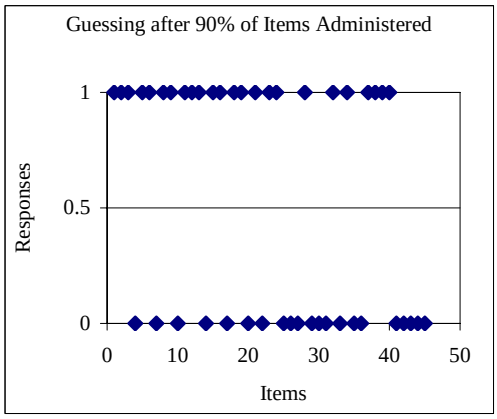


Responses for a High Ability Examinee for Various Guessing Behaviors

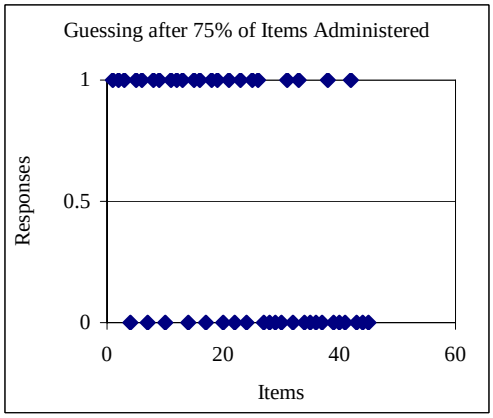
(44)



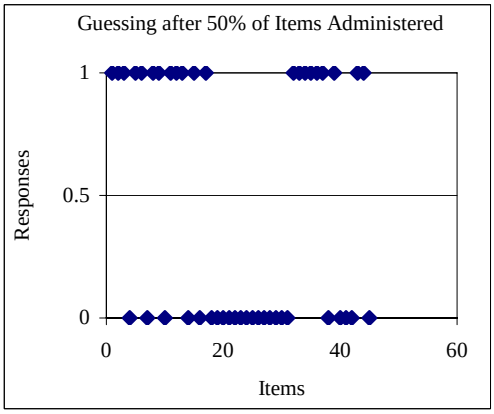
(45)



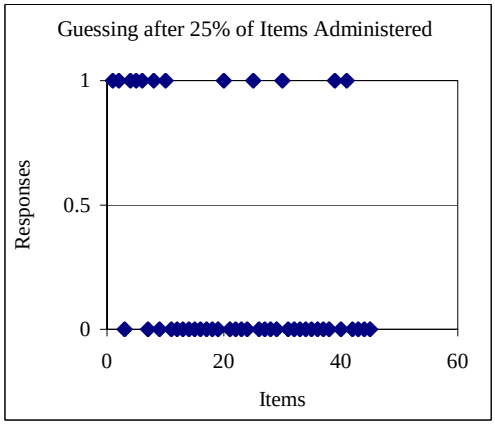
(46)



(47)

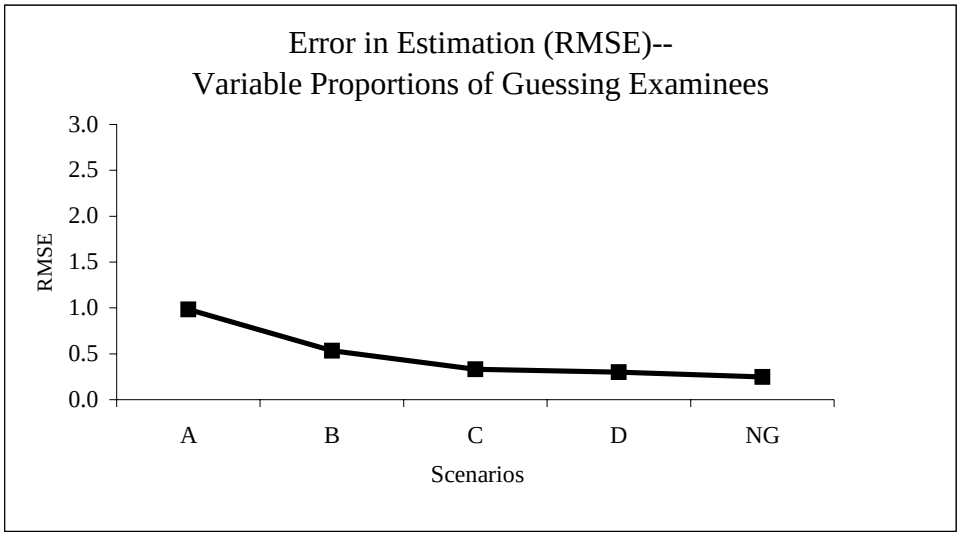


(48)



Error in Estimation due to Variable Proportions of Guessing Examinees

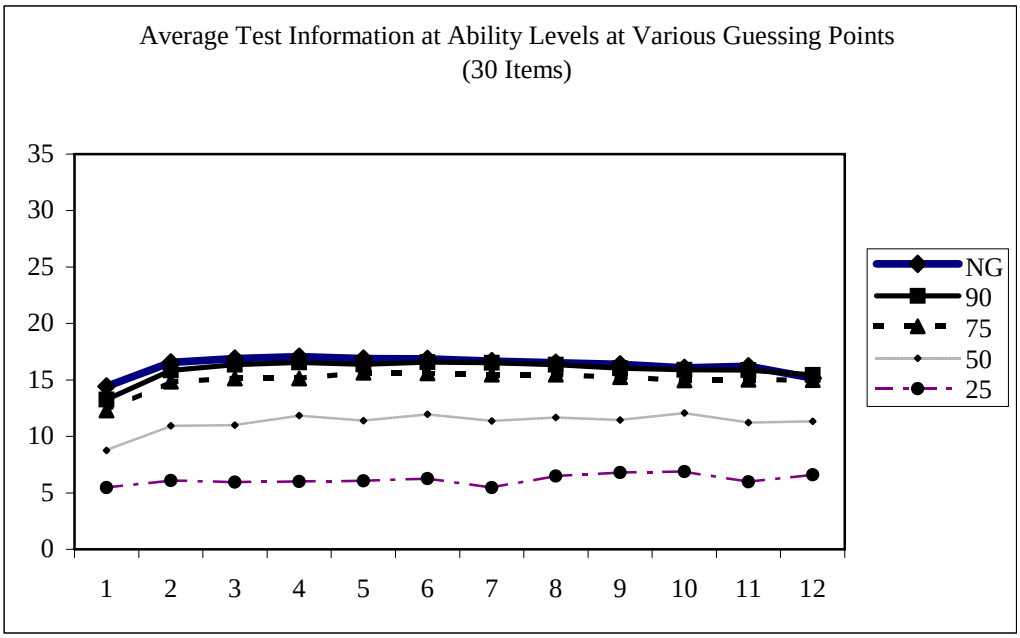
(49)



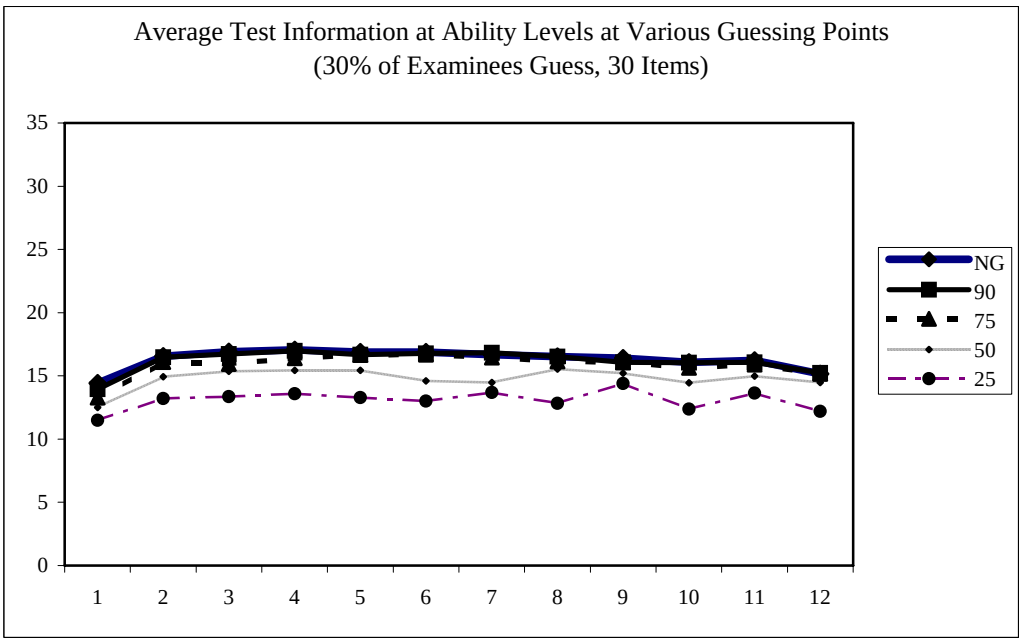
Note: Guessing introduced for 60% of examinees (flagged) at each ability level
Scenario A: 20% of flagged examinees begin guessing after 25% of items had been administered 80% guess towards the end (after 90% of items have been administered)
Scenario B: 20% of flagged examinees begin guessing after 50% of items had been administered 80% guess towards the end
Scenario C: 20% of flagged examinees begin guessing after 75% of items had been administered 80% guess towards the end
Scenario D: All flagged examinees begin guessing towards the end (after 90% of items)

Average Test Information at Ability Levels at Various Guessing Points
(Guessing Introduced for All vs. 30% of Examinees)

(50)



(51)

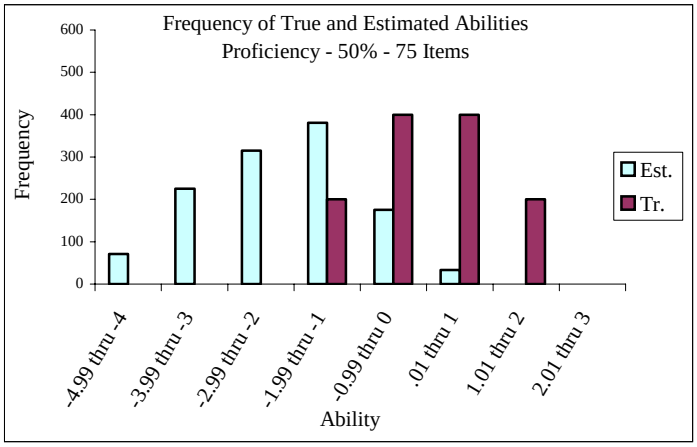


APPENDIX C

Distribution of Examinees in True and Estimated Ability Intervals

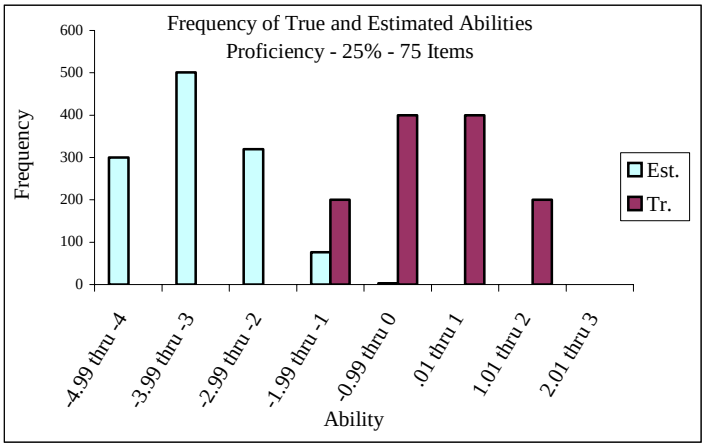
(55)

Ability	Est.	Tr.
-4.99 thru -4	71	0
-3.99 thru -3	225	0
-2.99 thru -2	315	0
-1.99 thru -1	381	200
-0.99 thru 0	175	400
.01 thru 1	33	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



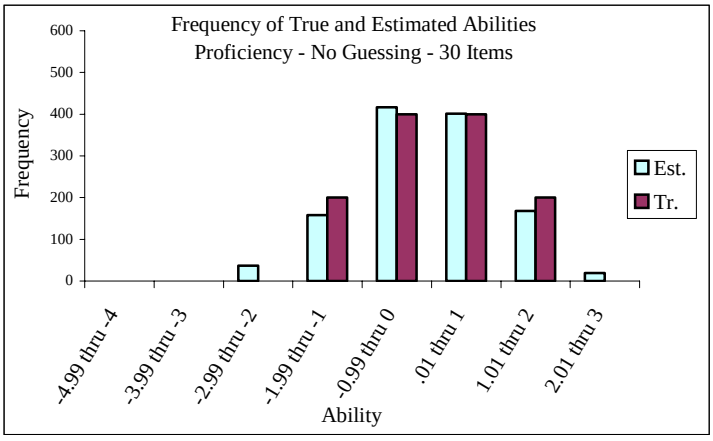
(56)

Ability	Est.	Tr.
-4.99 thru -4	300	0
-3.99 thru -3	501	0
-2.99 thru -2	320	0
-1.99 thru -1	76	200
-0.99 thru 0	3	400
.01 thru 1	0	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



(57)

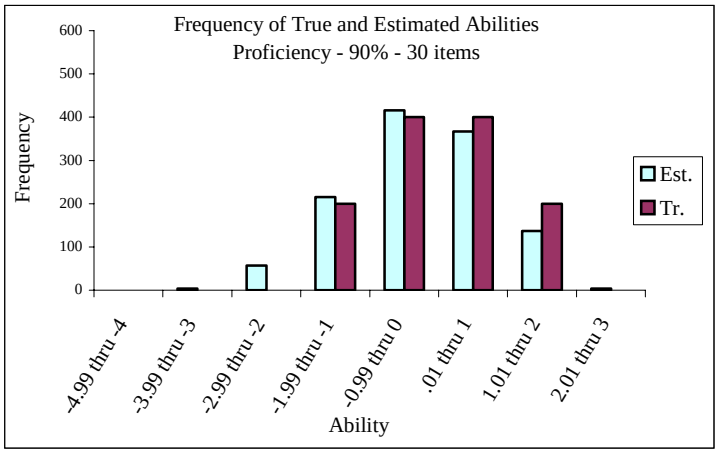
Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	0	0
-2.99 thru -2	37	0
-1.99 thru -1	158	200
-0.99 thru 0	417	400
.01 thru 1	401	400
1.01 thru 2	168	200
2.01 thru 3	19	0
Total	1200	0



Distribution of Examinees in True and Estimated Ability Intervals

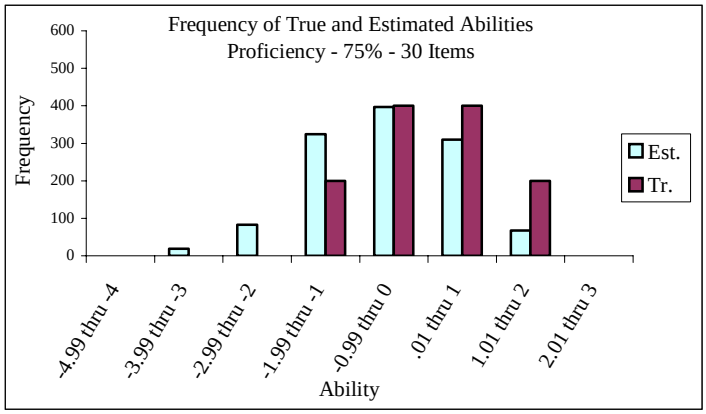
(58)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	4	0
-2.99 thru -2	57	0
-1.99 thru -1	215	200
-0.99 thru 0	416	400
.01 thru 1	367	400
1.01 thru 2	137	200
2.01 thru 3	4	0
Total	1200	0



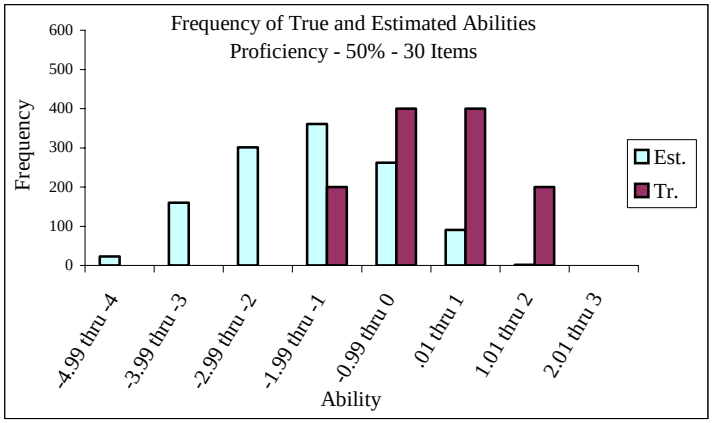
(59)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	19	0
-2.99 thru -2	83	0
-1.99 thru -1	324	200
-0.99 thru 0	397	400
.01 thru 1	310	400
1.01 thru 2	67	200
2.01 thru 3	0	0
Total	1200	0



(60)

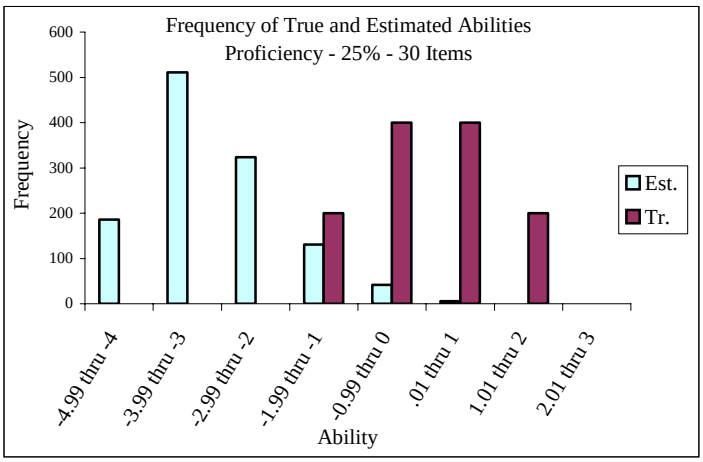
Ability	Est.	Tr.
-4.99 thru -4	23	0
-3.99 thru -3	160	0
-2.99 thru -2	301	0
-1.99 thru -1	361	200
-0.99 thru 0	262	400
.01 thru 1	91	400
1.01 thru 2	2	200
2.01 thru 3	0	0
Total	1200	0



Distribution of Examinees in True and Estimated Ability Intervals

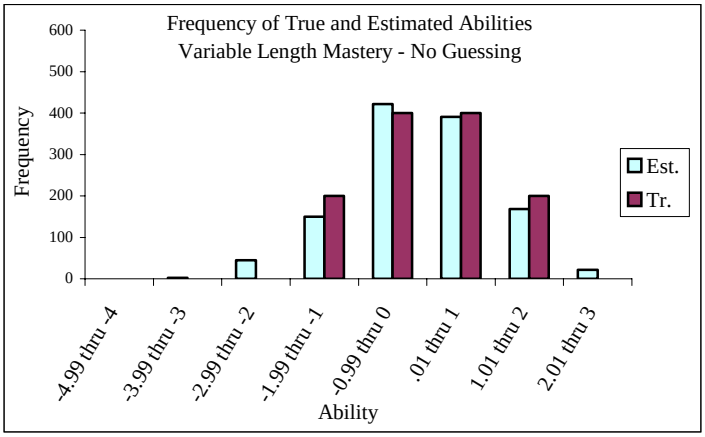
(61)

Ability	Est.	Tr.
-4.99 thru -4	186	0
-3.99 thru -3	511	0
-2.99 thru -2	324	0
-1.99 thru -1	131	200
-0.99 thru 0	42	400
.01 thru 1	6	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0



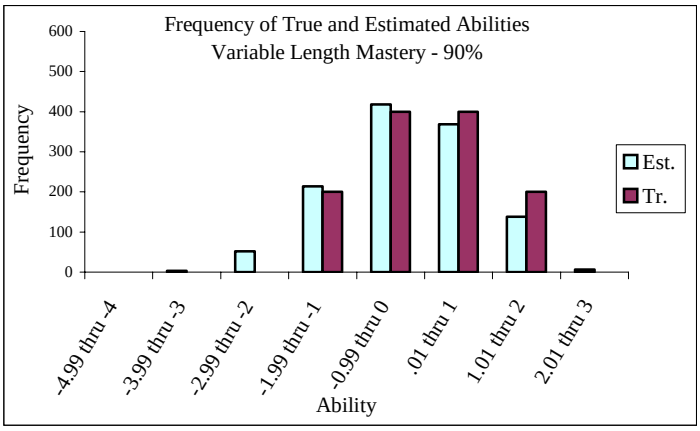
(62)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	2	0
-2.99 thru -2	45	0
-1.99 thru -1	150	200
-0.99 thru 0	422	400
.01 thru 1	391	400
1.01 thru 2	168	200
2.01 thru 3	22	0
Total	1200	0



(63)

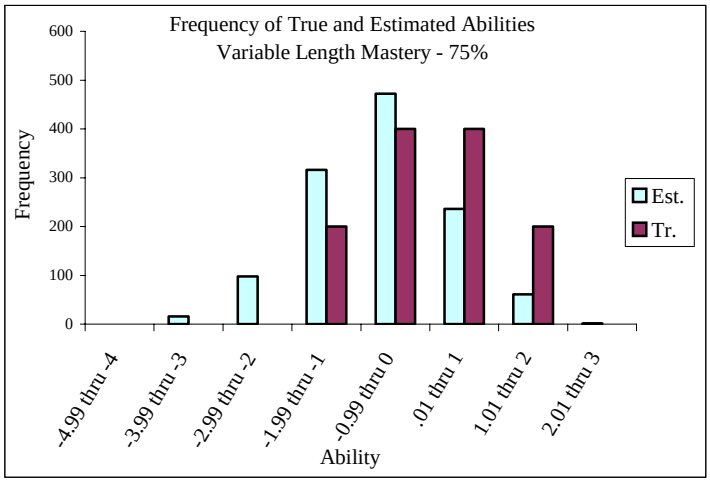
Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	3	0
-2.99 thru -2	52	0
-1.99 thru -1	214	200
-0.99 thru 0	418	400
.01 thru 1	369	400
1.01 thru 2	138	200
2.01 thru 3	6	0
Total	1200	0



Distribution of Examinees in True and Estimated Ability Intervals

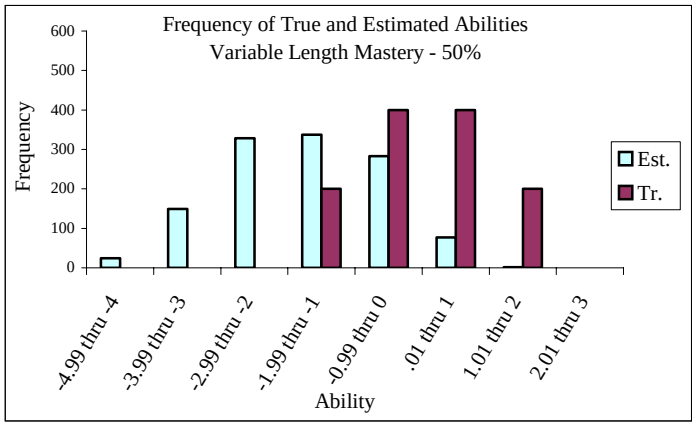
(64)

Ability	Est.	Tr.
-4.99 thru -4	0	0
-3.99 thru -3	16	0
-2.99 thru -2	98	0
-1.99 thru -1	316	200
-0.99 thru 0	472	400
.01 thru 1	236	400
1.01 thru 2	61	200
2.01 thru 3	1	0
Total	1200	0



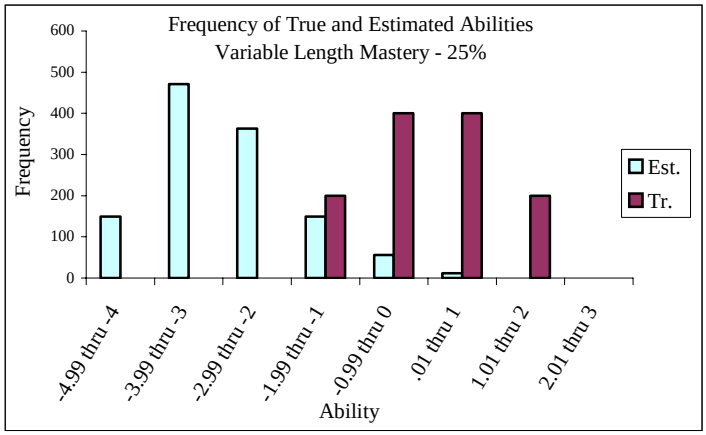
(65)

Ability	Est.	Tr.
-4.99 thru -4	24	0
-3.99 thru -3	149	0
-2.99 thru -2	328	0
-1.99 thru -1	337	200
-0.99 thru 0	283	400
.01 thru 1	77	400
1.01 thru 2	2	200
2.01 thru 3	0	0
Total	1200	0



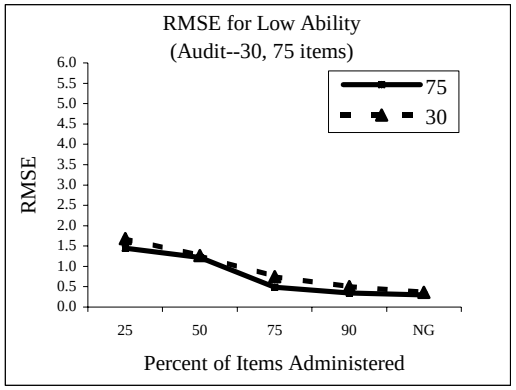
(66)

Ability	Est.	Tr.
-4.99 thru -4	149	0
-3.99 thru -3	471	0
-2.99 thru -2	363	0
-1.99 thru -1	149	200
-0.99 thru 0	56	400
.01 thru 1	12	400
1.01 thru 2	0	200
2.01 thru 3	0	0
Total	1200	0

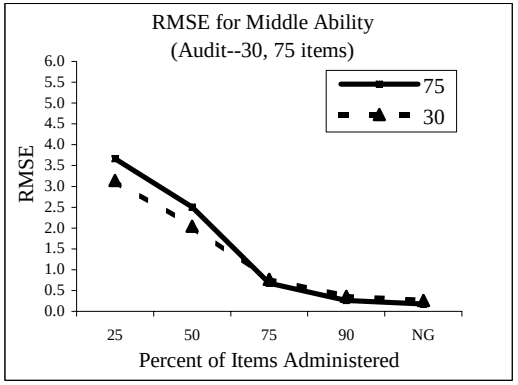


RMSE of Estimates around True Ability for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

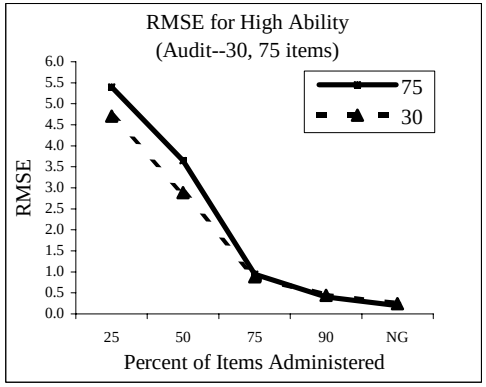
(67)



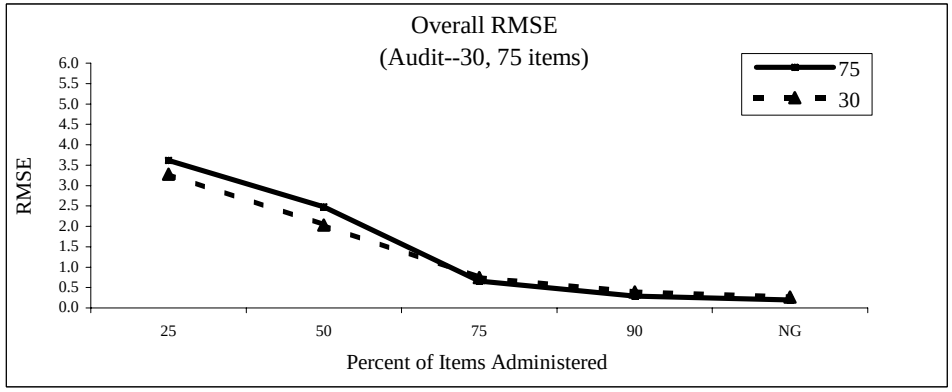
(68)



(69)

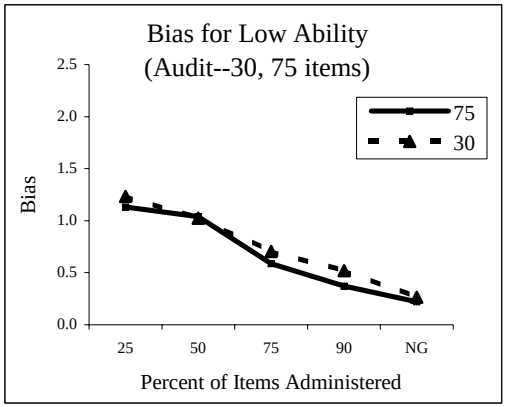


(70)

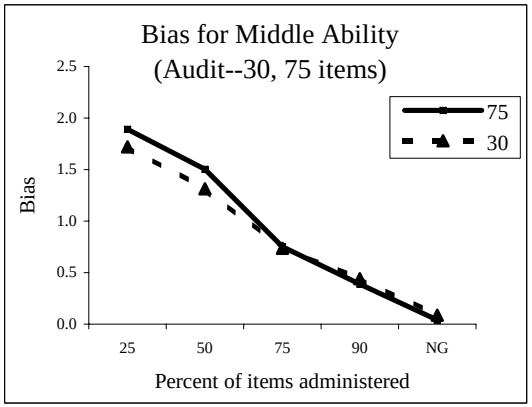


Bias in Estimates around True ability for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

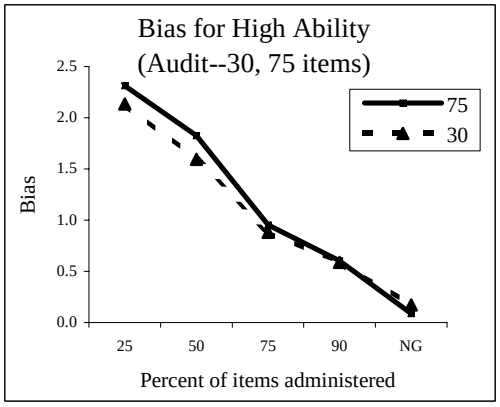
(71)



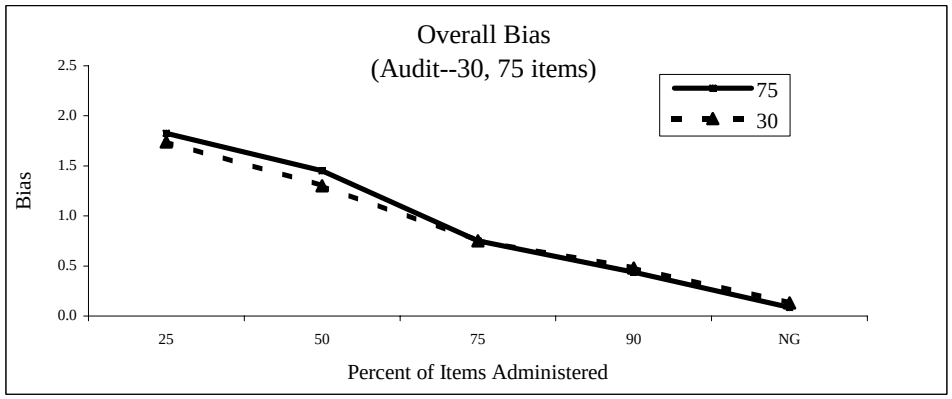
(72)



(73)

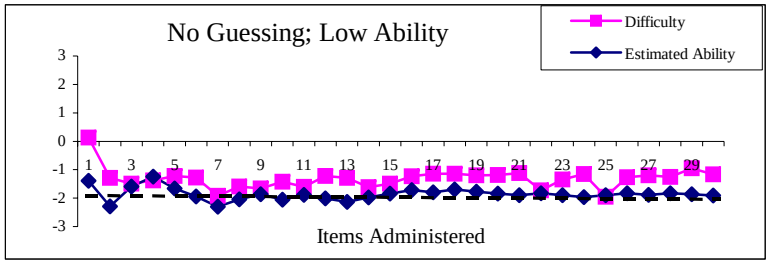


(74)

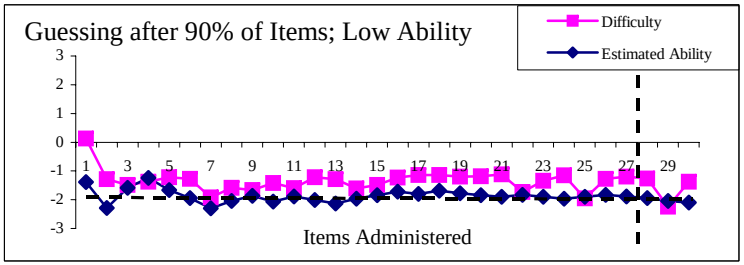


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)

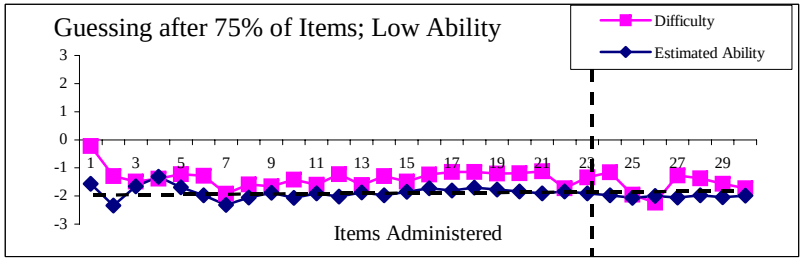
(75)



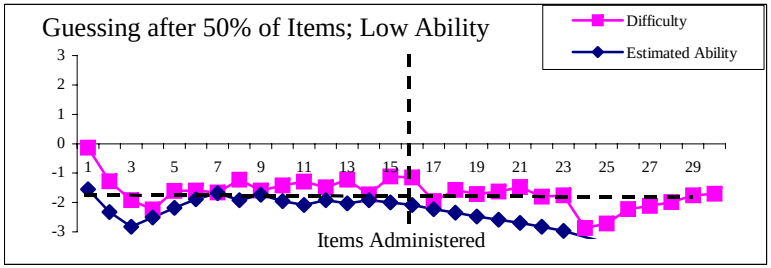
(76)



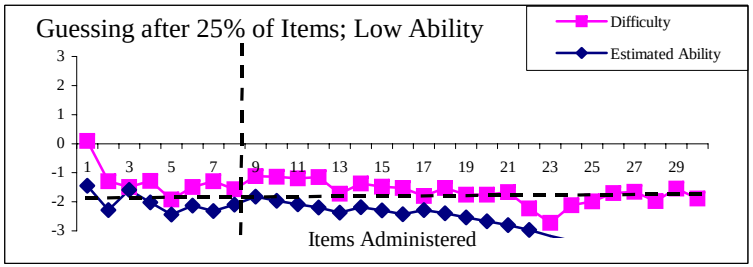
(77)



(78)

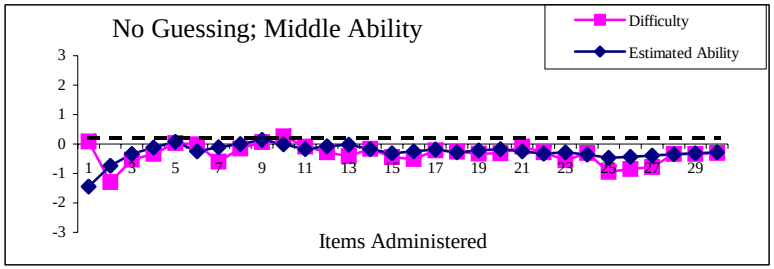


(79)

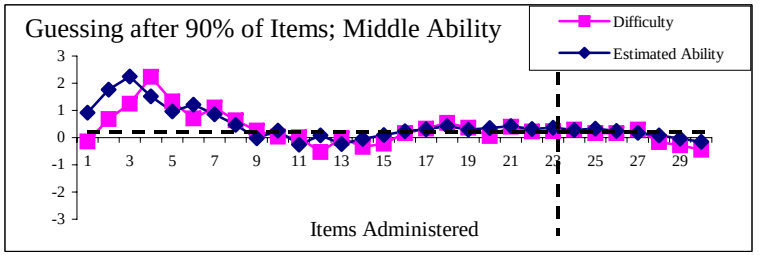


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)

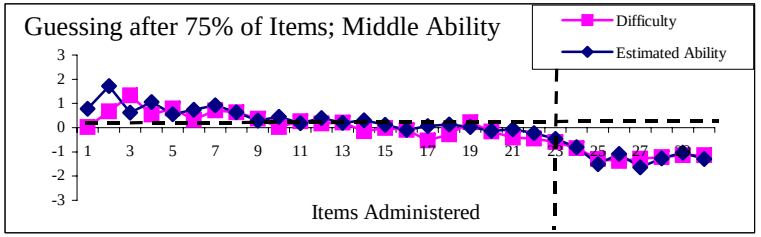
(80)



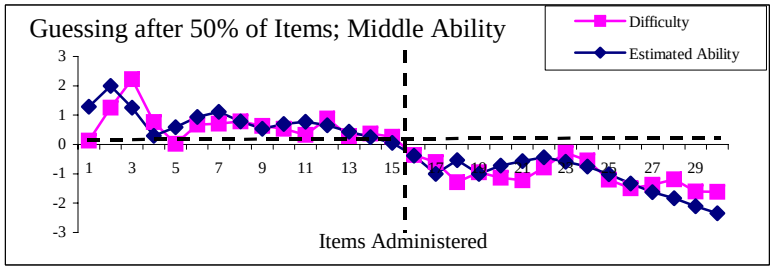
(81)



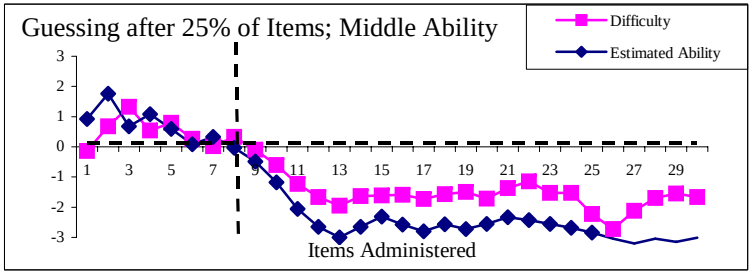
(82)



(83)

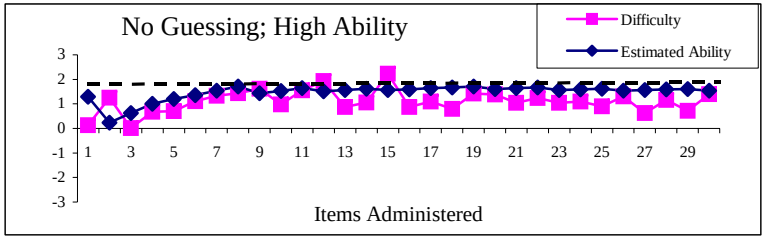


(84)

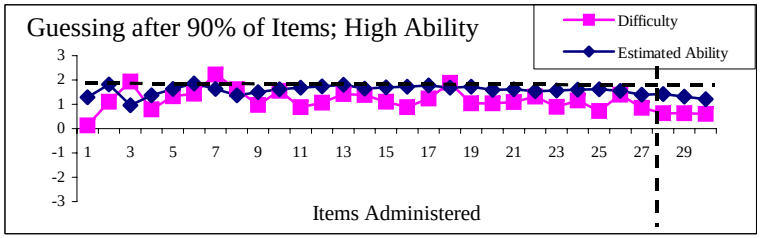


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--30 Items)

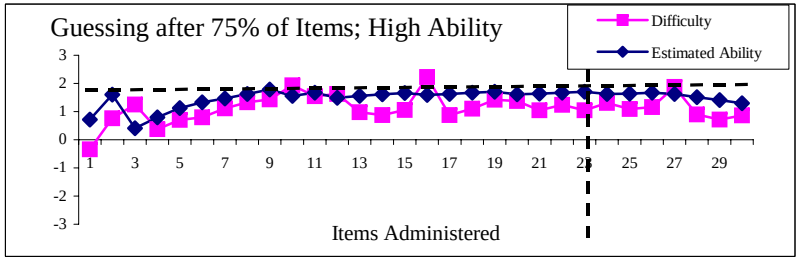
(85)



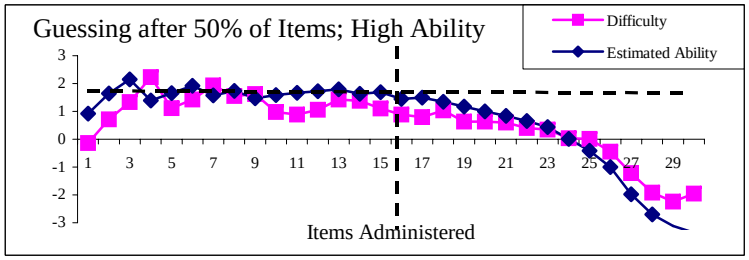
(86)



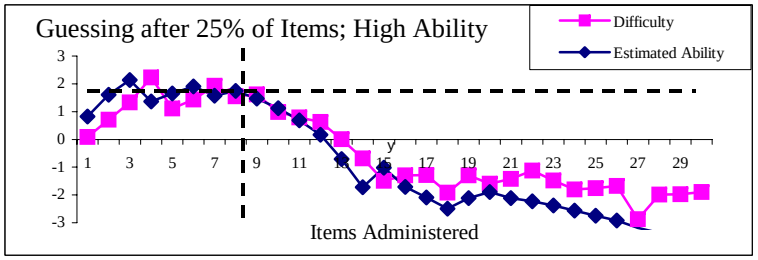
(87)



(88)

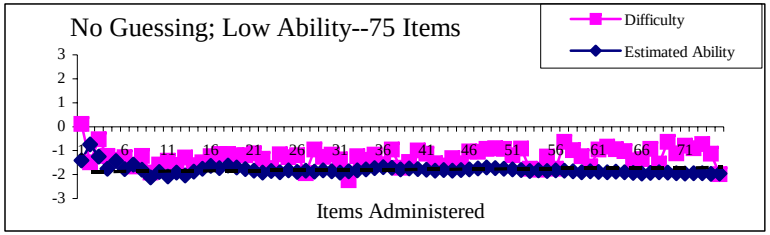


(89)

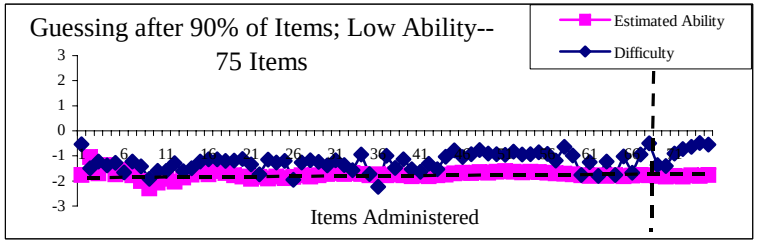


CAT Administration for a Low Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

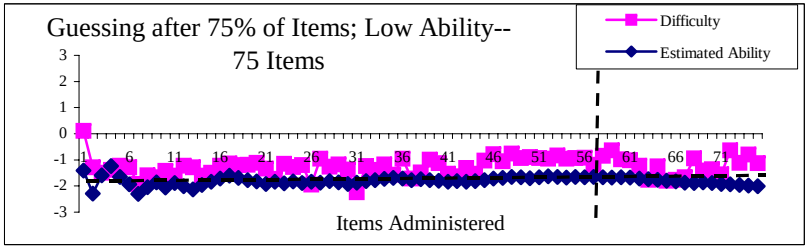
(90)



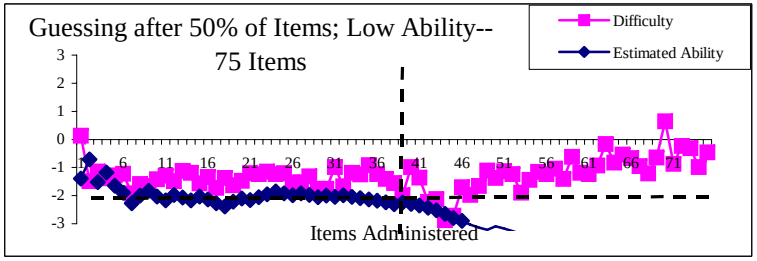
(91)



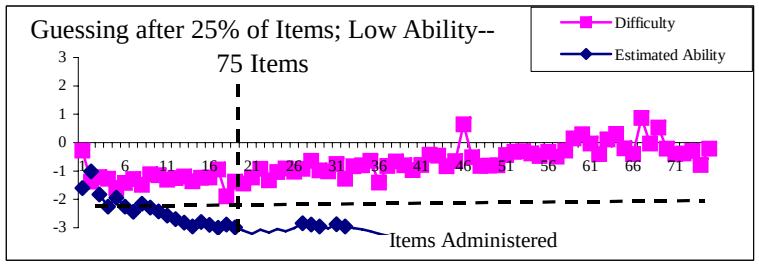
(92)



(93)

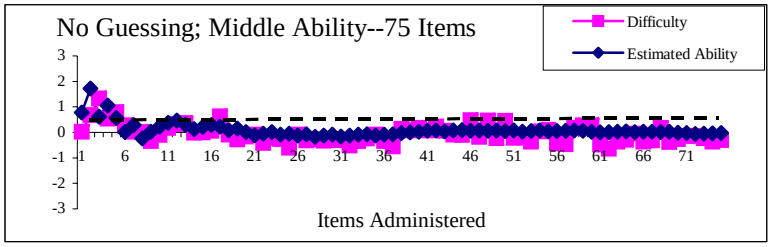


(94)

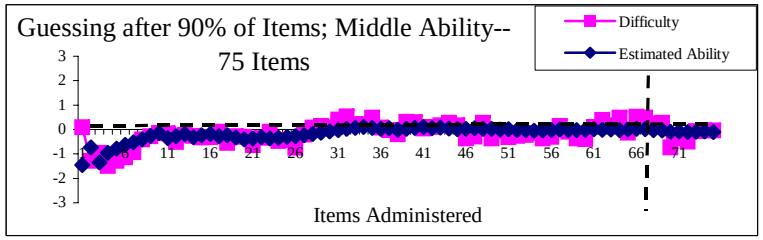


CAT Administration for a Middle Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

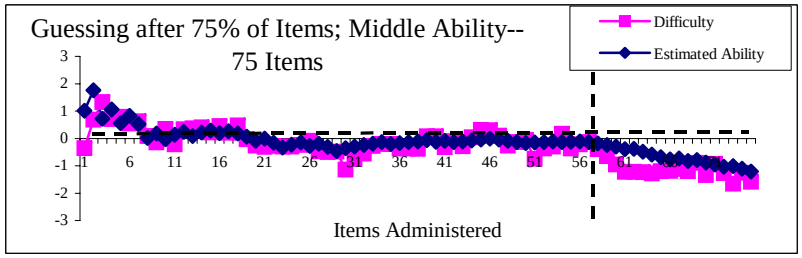
(95)



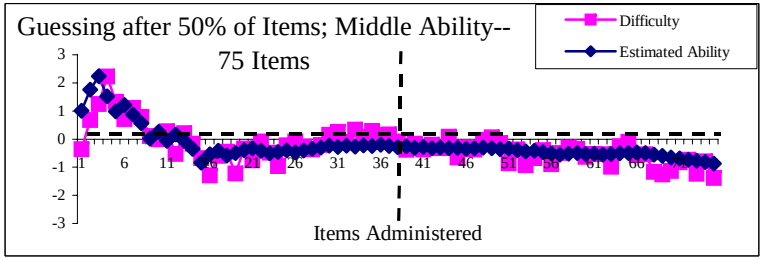
(96)



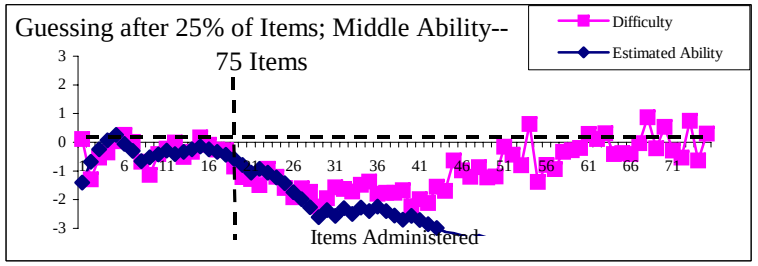
(97)



(98)

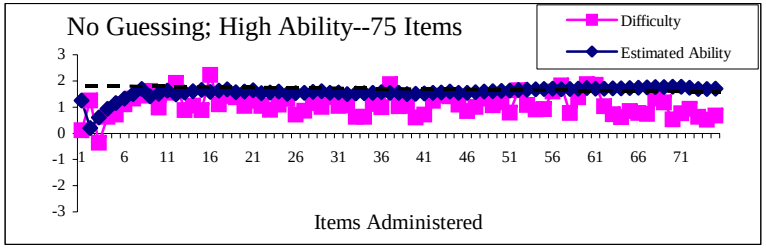


(99)

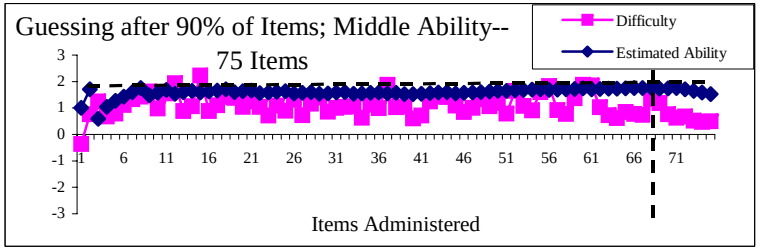


CAT Administration for a High Ability Examinee who Guesses at a Certain Point in the Test (AUDIT--75 Items)

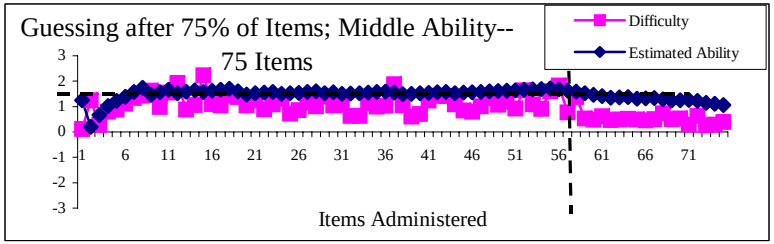
(100)



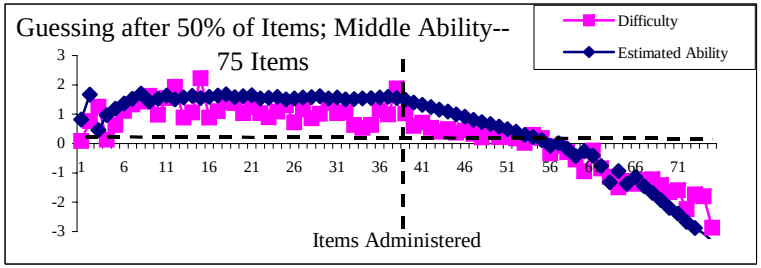
(101)



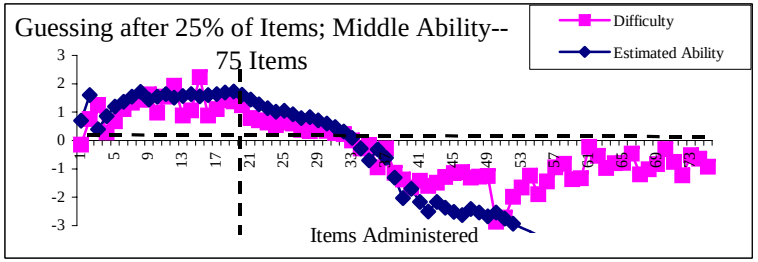
(102)



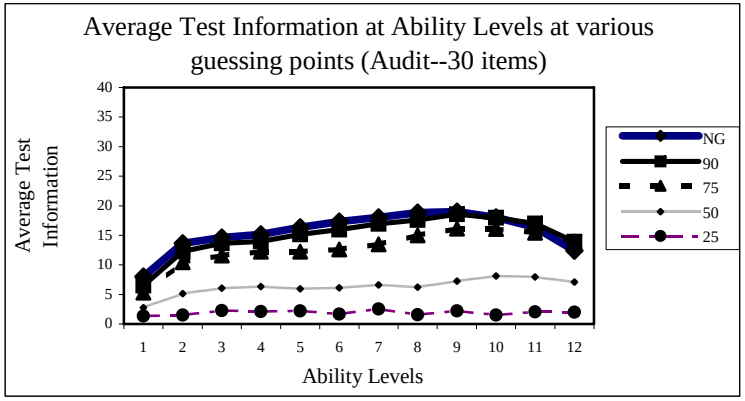
(103)



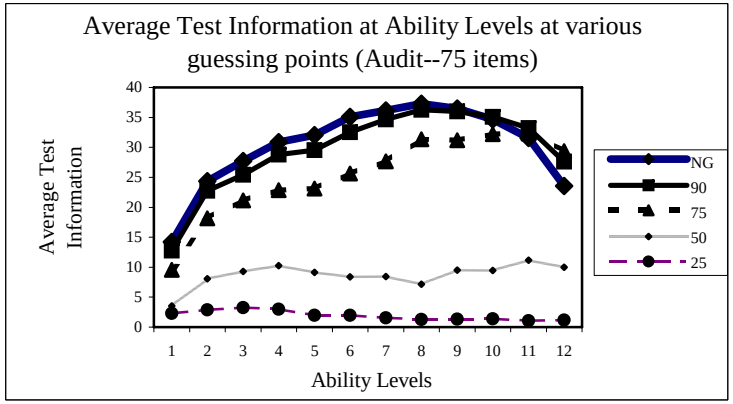
(104)



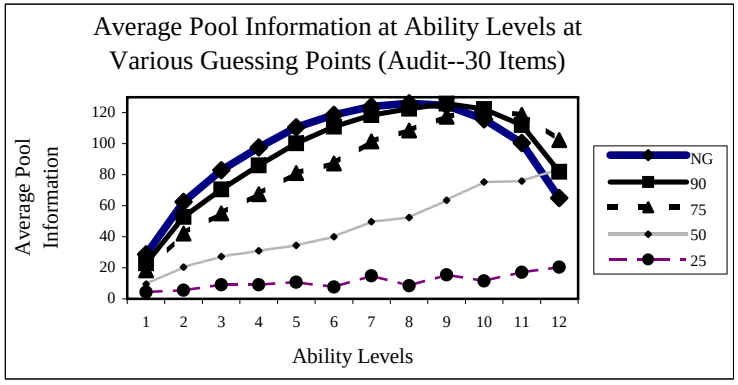
(105)



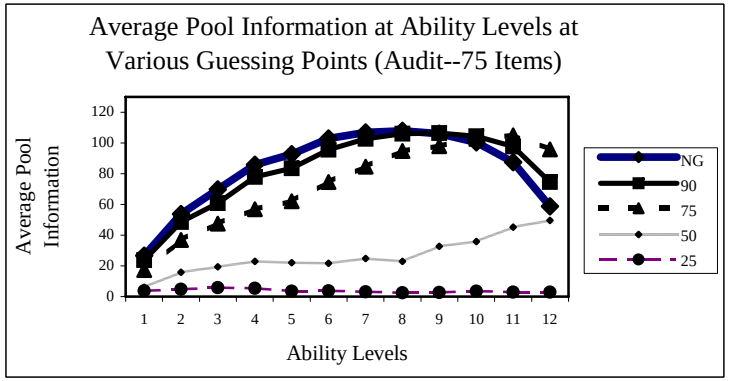
(106)



(107)

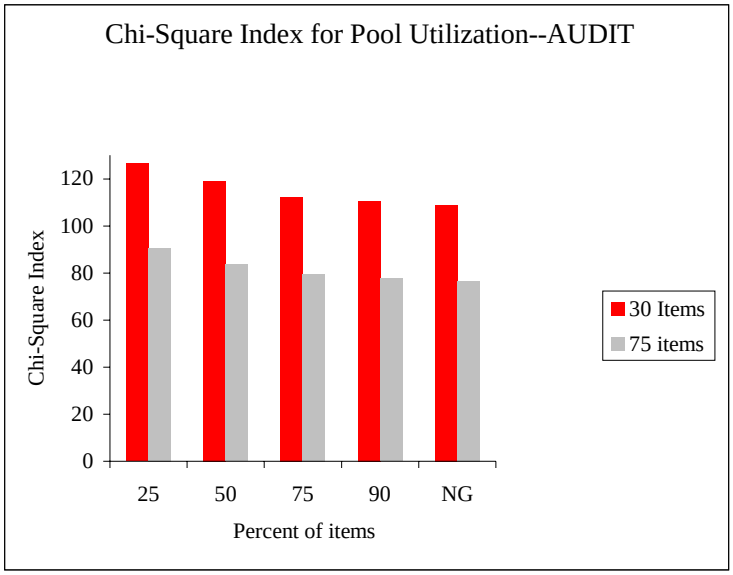


(108)



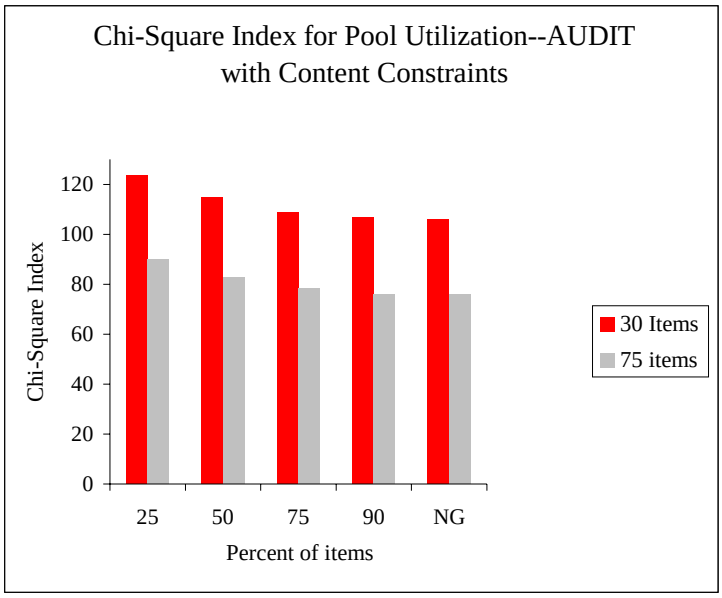
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 and 75 items on AUDIT without/with Content Constraints)

(109)



	30 items	75 items
25	126.48	90.74
50	118.98	83.67
75	112.29	79.55
90	110.51	77.59
NG	108.90	76.67

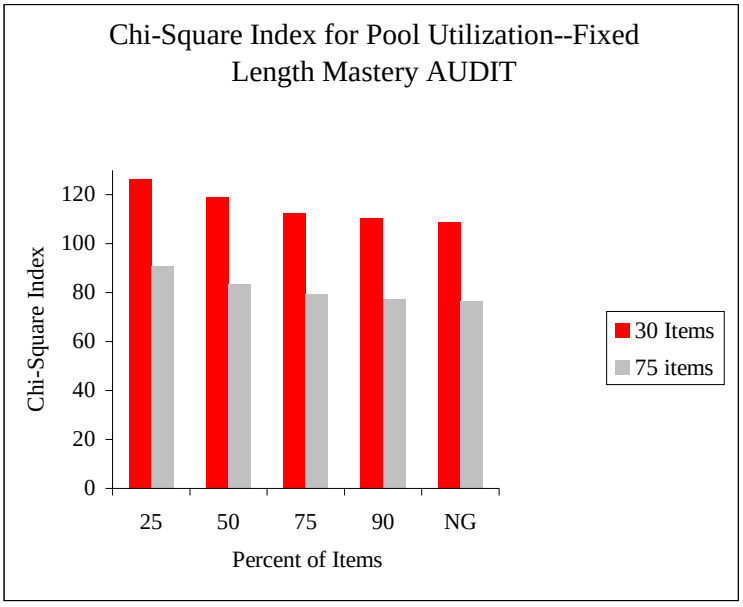
(110)



	30 items	75 items
25	123.71	89.99
50	114.81	82.55
75	108.70	78.26
90	106.77	75.92
NG	105.92	75.79

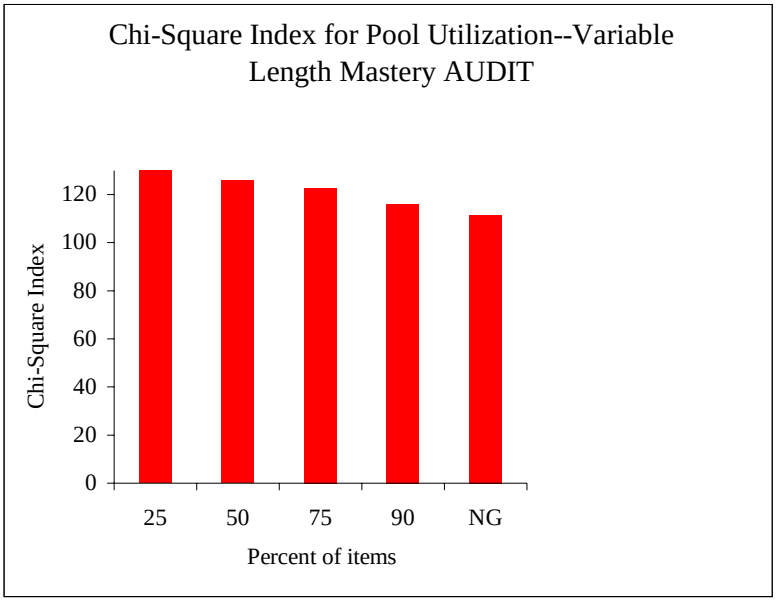
Pool Utilization Index at each Ability Level for 5 Guessing Scenarios
(Mastery Testing with AICPA Parameters for Fixed/Variable Length AUDIT)

(111)



	30 items	75 items
25	126.48	90.74
50	118.98	83.67
75	112.29	79.55
90	110.51	77.59
NG	108.90	76.67

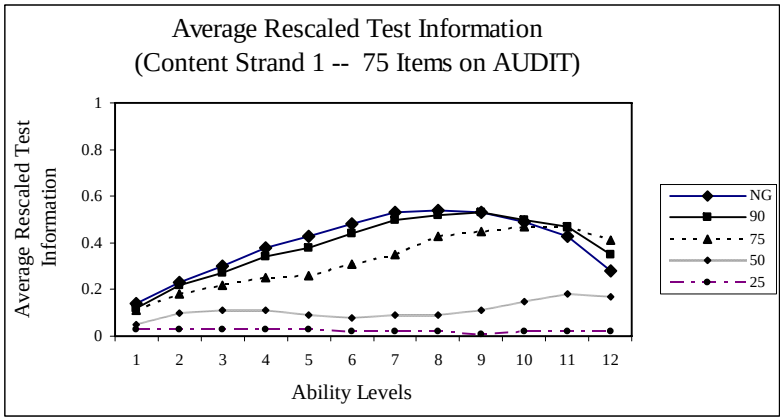
(112)



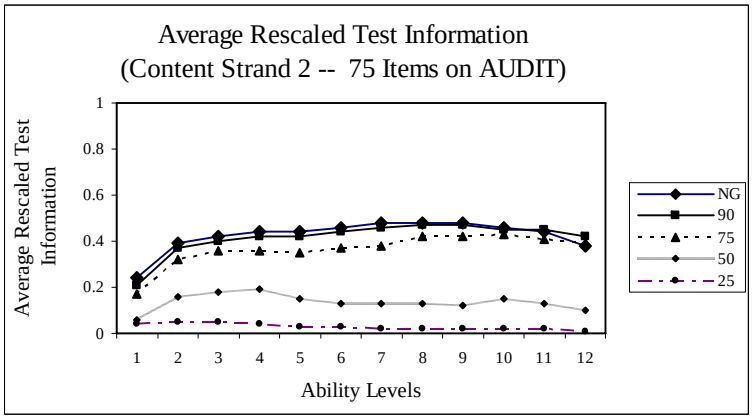
Variable Length	
25	129.98
50	125.85
75	122.66
90	116.10
NG	111.52

Average Rescaled Test Information in each Content Area at each Ability Level
(Performance Testing with AICPA Parameters for 75 Items on AUDIT)

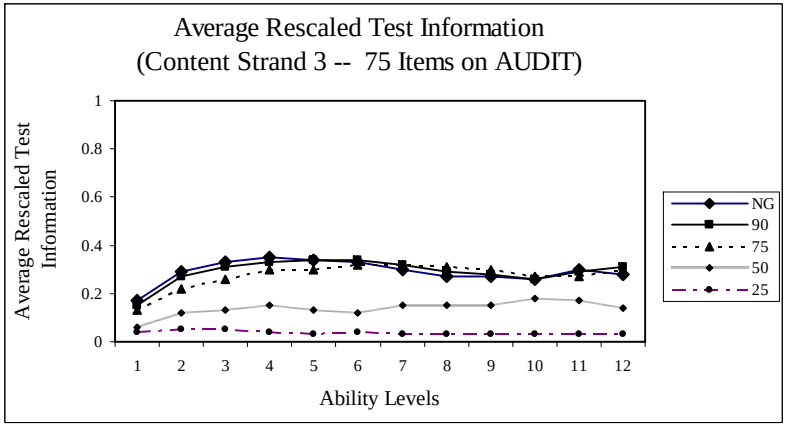
(113)



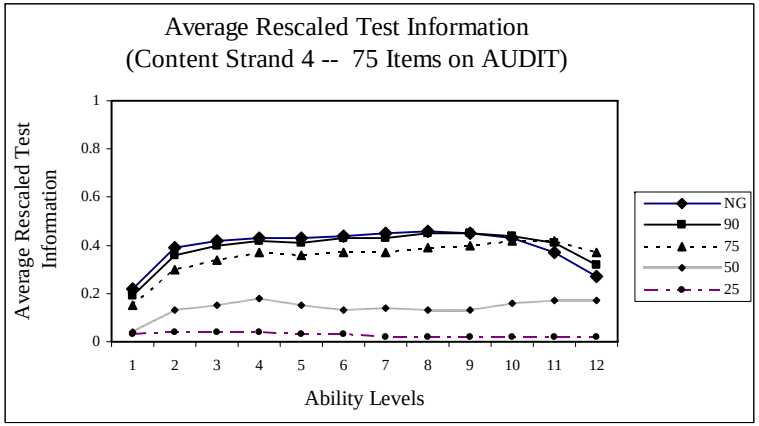
(114)



(115)

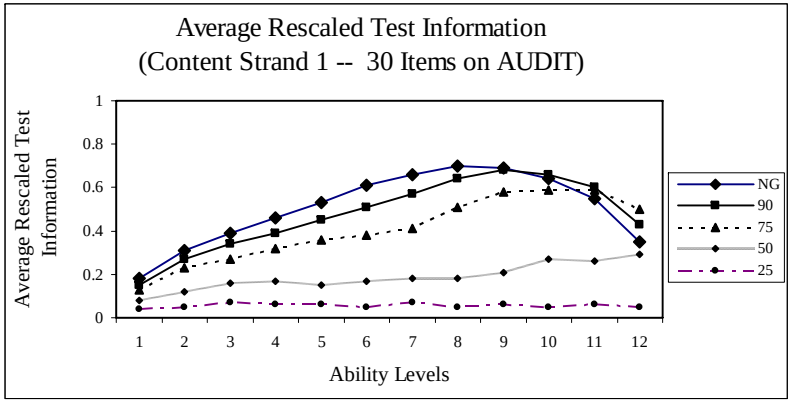


(116)

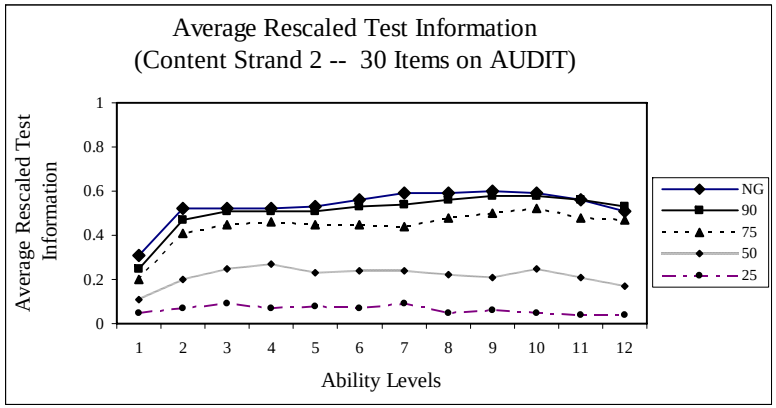


Average Rescaled Test Information in each Content Area at each Ability Level
(Performance Testing with AICPA Parameters for 30 Items on AUDIT)

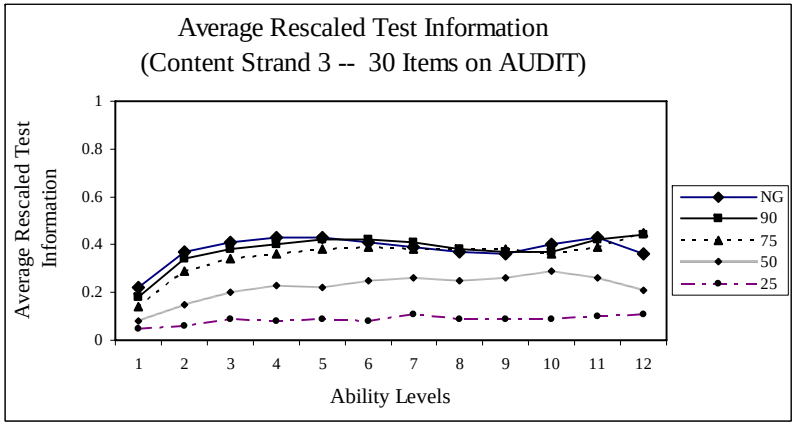
(117)



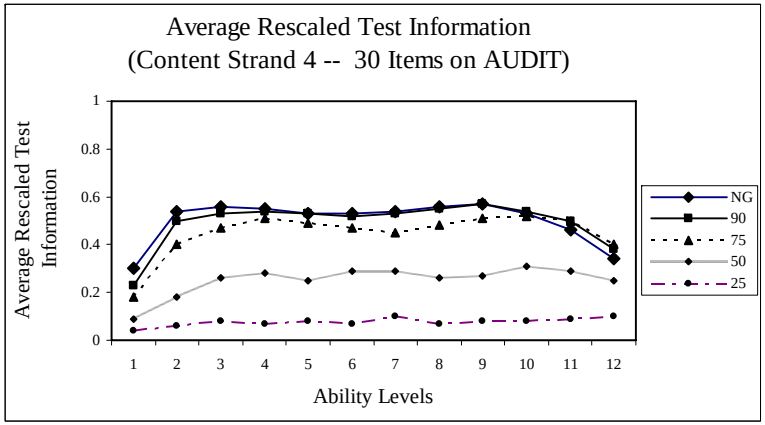
(118)



(119)

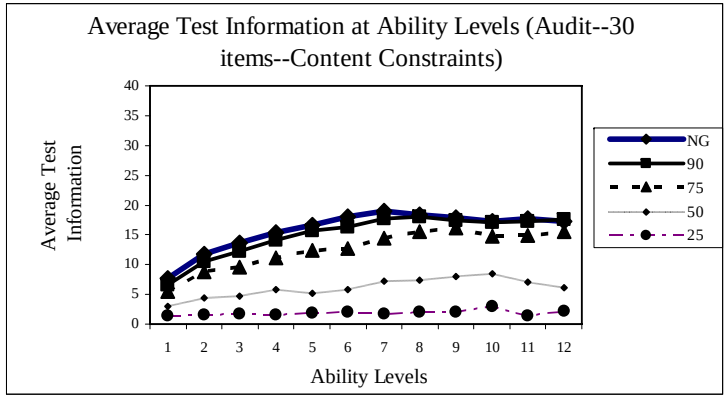


(120)

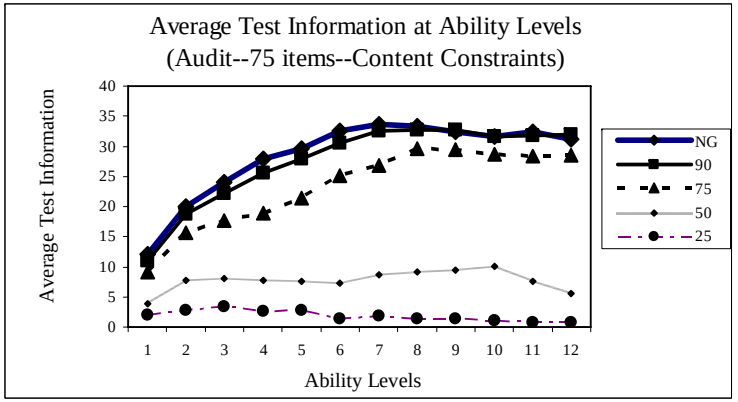


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints)

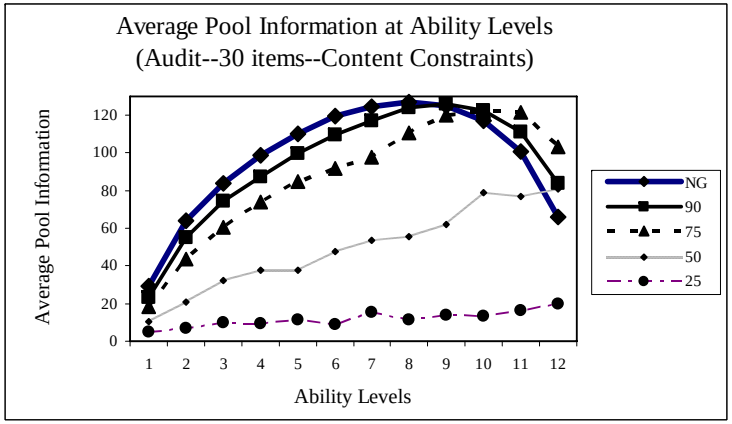
(121)



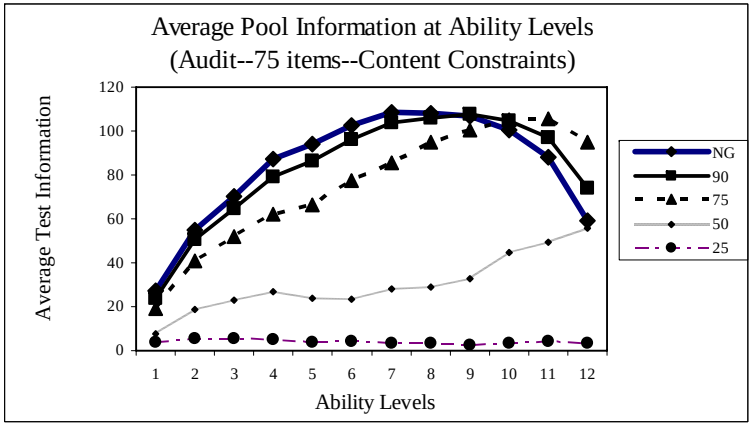
(122)



(123)

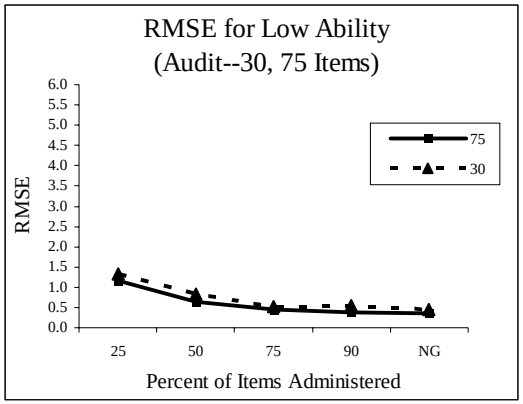


(124)

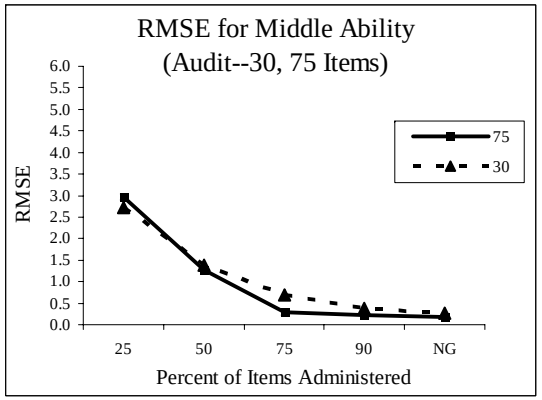


RMSE of Estimates around True Ability for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints
b-Parameter Increased by a Constant for Content Strand 1)

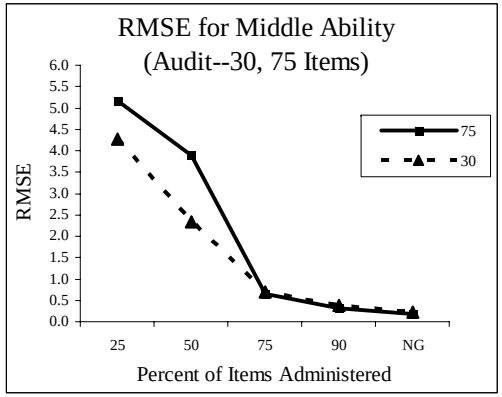
(125)



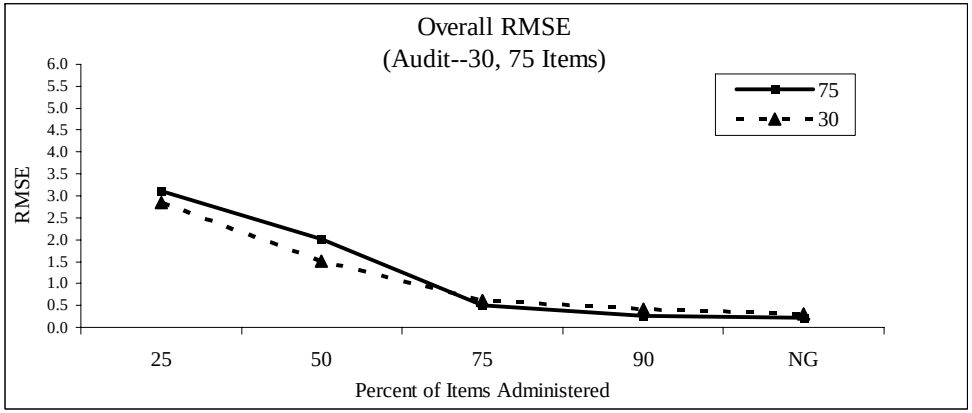
(126)



(127)

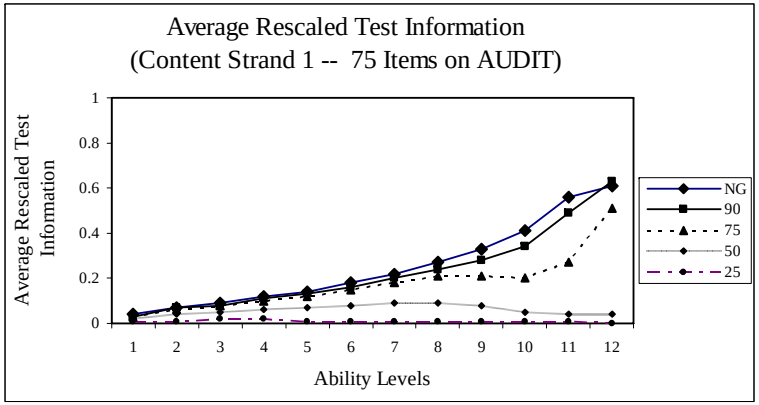


(128)

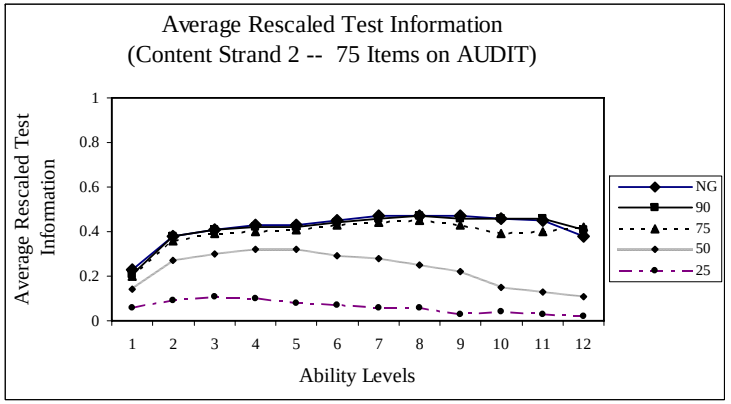


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 75 Items on AUDIT with Content Constraints)
b-Parameter Increased by a Constant for Content Strand 1)

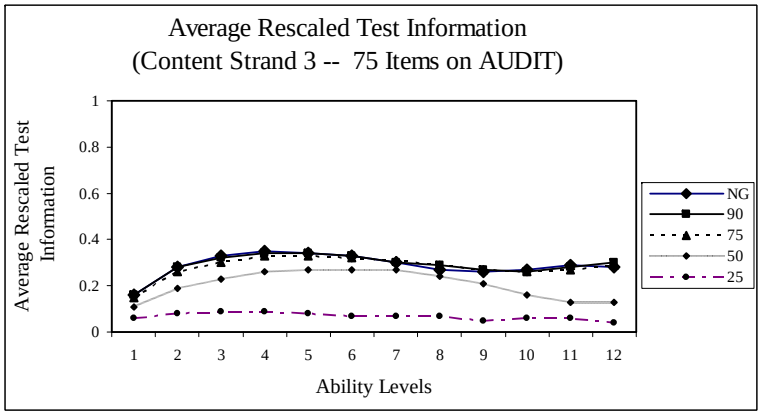
(129)



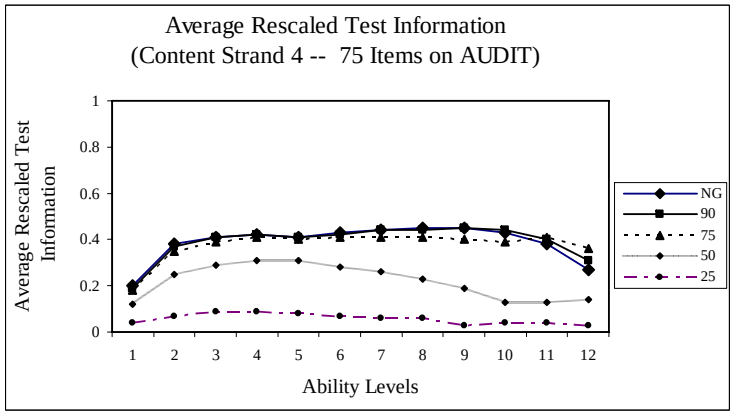
(130)



(131)

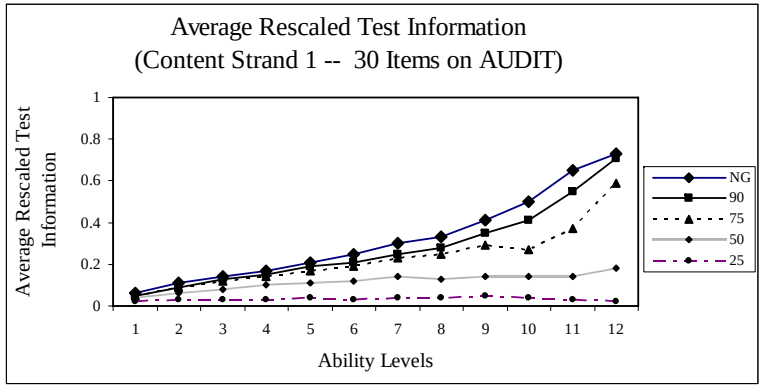


(132)

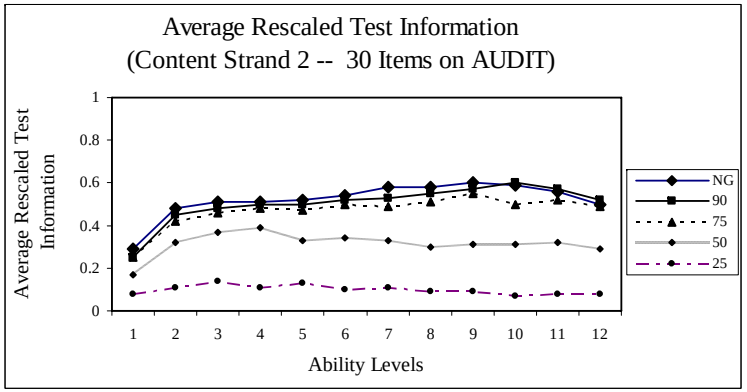


Average Rescaled Test Information at each Ability Level for 5 Guessing Scenarios
(Performance Testing with AICPA Parameters for 30 Items on AUDIT with Content Constraints)
b-Parameter Increased by a Constant for Content Strand 1)

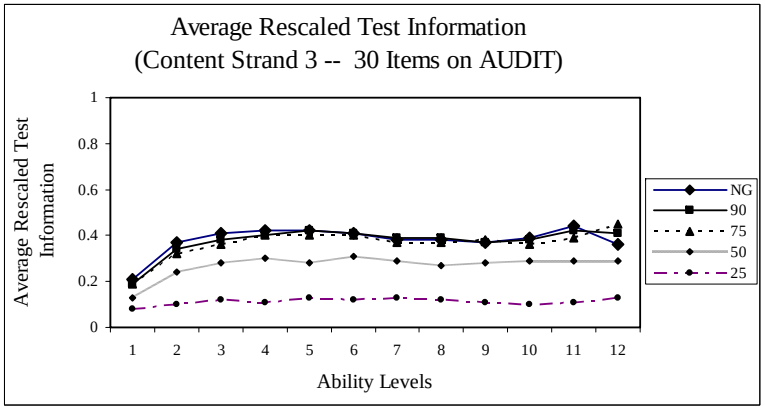
(133)



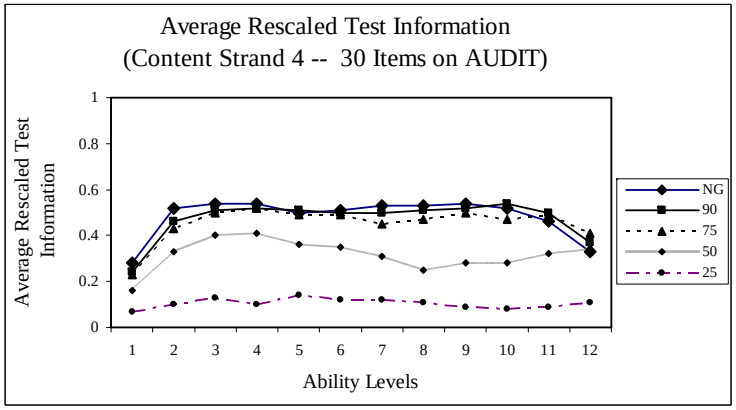
(134)



(135)

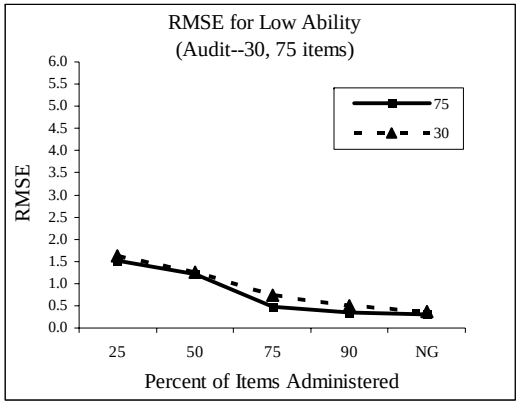


(136)

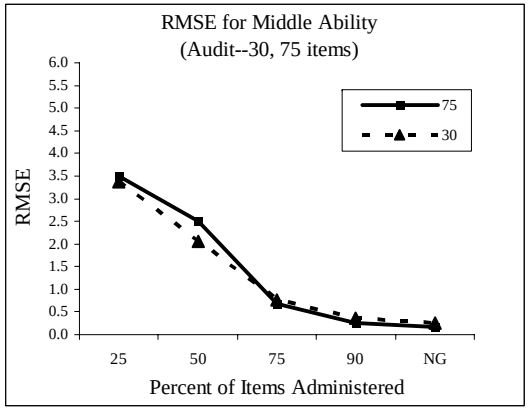


RMSE of Estimates around True Ability for 5 Guessing Scenarios
(Fixed Length Mastery Testing with AICPA Parameters for 30 and 75 Items on AUDIT with Content Constraints)

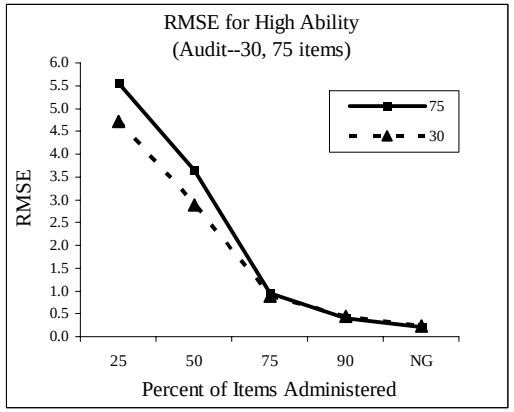
(137)



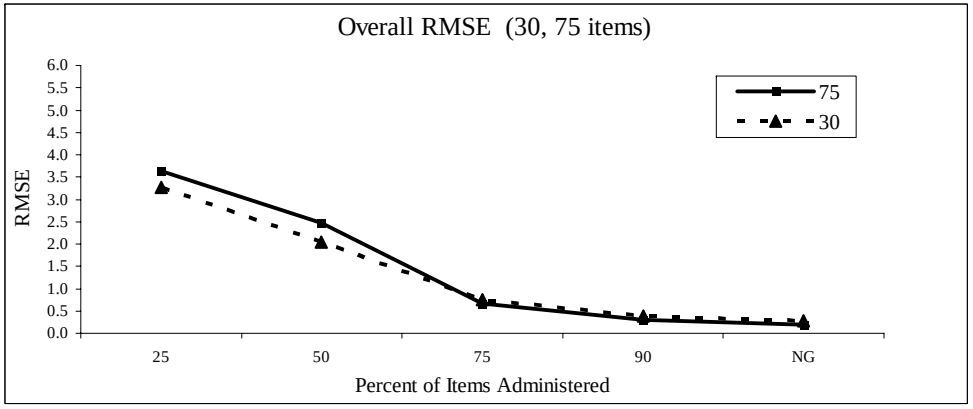
(138)



(139)

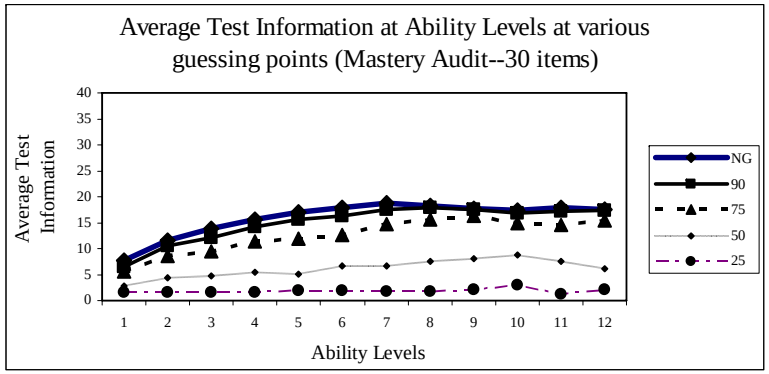


(140)

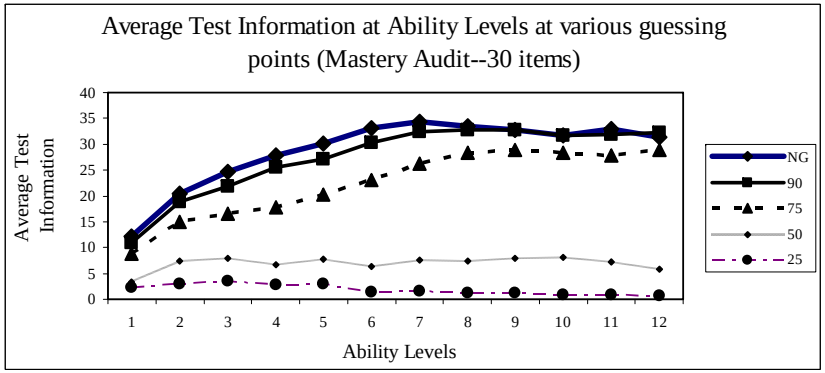


Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios
(Fixed Length Mastery Testing with AICPA Parameters for 30 and 75 Items on AUDIT)

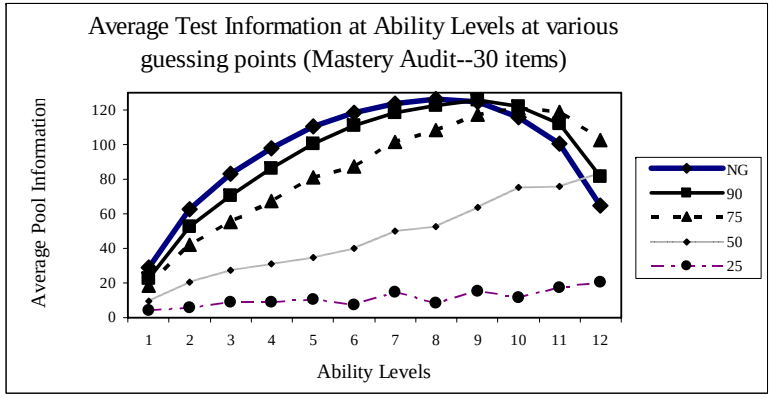
(141)



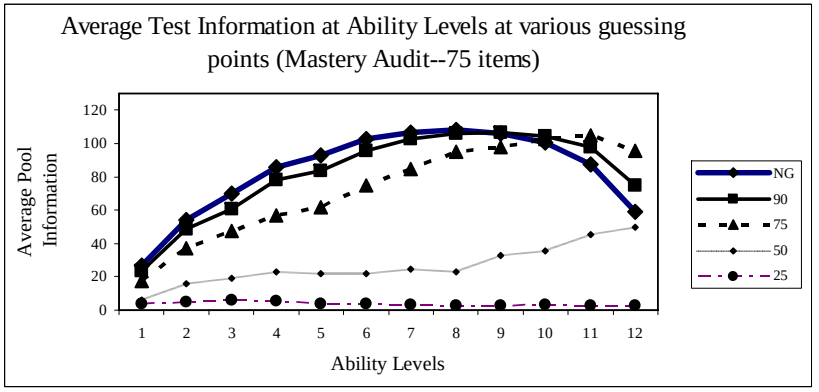
(142)



(143)

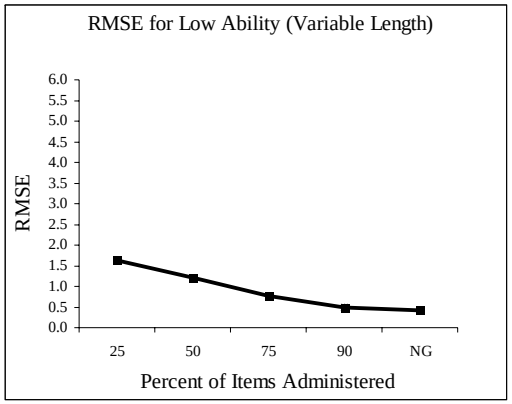


(144)

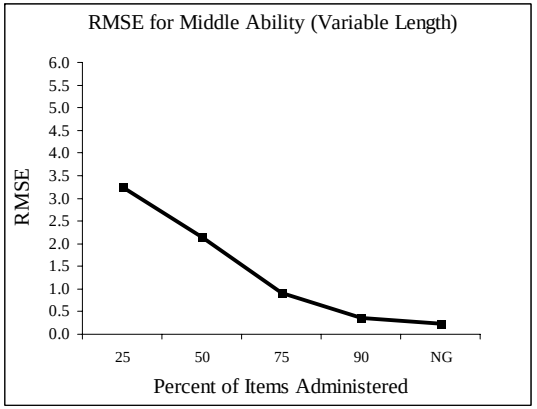


RMSE of estimates around true ability for 5 guessing scenarios
(Variable Length Mastery testing with AICPA parameters for AUDIT)

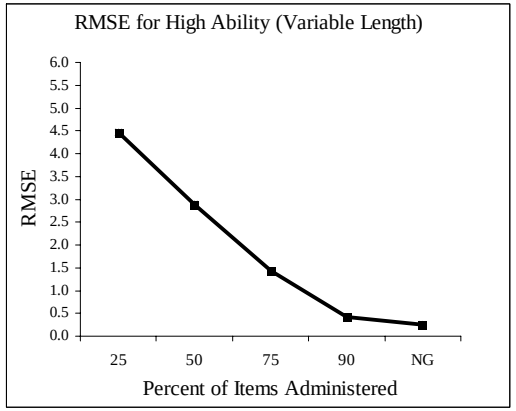
(145)



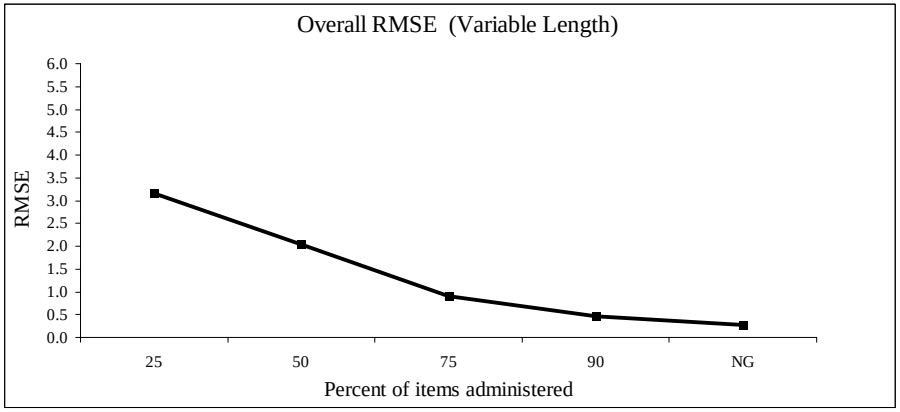
(146)



(147)

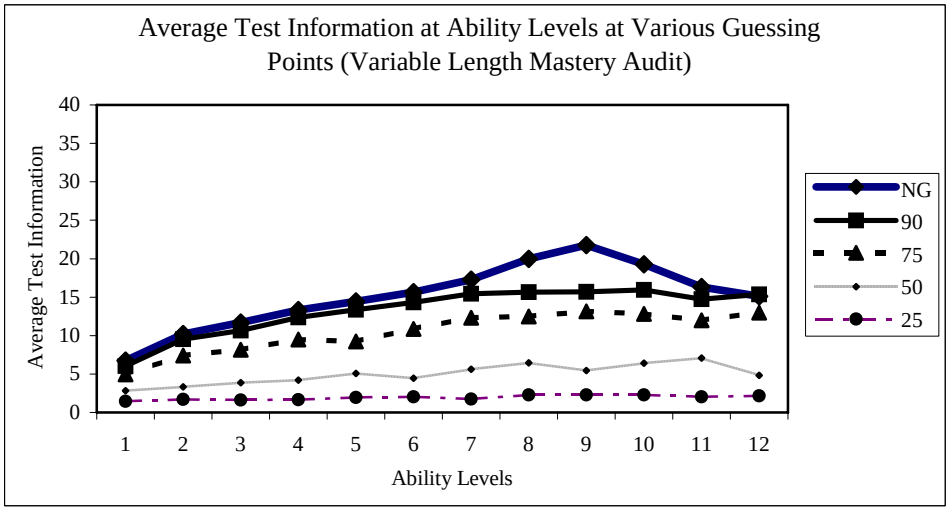


(148)



Average Test and Pool Information at each Ability Level for 5 Guessing Scenarios
(Variable Length Mastery Testing with AICPA Parameters -- AUDIT)

(149)



(150)