Comparison of Multi-stage Tests with Computerized Adaptive and Paper and Pencil Tests

Ourania Rotou
Liane Patsula
Steffen Manfred
Saba Rizavi

Educational Testing Service

Paper presented at the annual meeting of the American Educational Research Association (AERA) and the National Council on Measurement in Education (NCME) held between April 21 to 25, 2003, in Chicago, IL.

Copyright © 2003 by Educational Testing Service. All Rights Reserved.

Abstract

Traditionally, the fixed length linear paper and pencil (P&P) has been the standard method of test delivery. However, with the advancement of technology, the popularity of administering tests using adaptive methods like computerized adaptive testing (CAT) and multi-stage testing (MST) has grown in the field of measurement in both theory and practice. In practice, there are several standardized tests that have sections that include only set-based items. To date, there is no study in the literature that compares these testing procedures when a test is completely set-based under various IRT models. This study investigates the measurement precision of MST to CAT and P&P tests for the 1-, 2-, and 3- PL models when the test is completely ser-based. Results showed that MST performed better for the 2- and 3-PL models than equivalent length P&P test in terms of reliability and conditional standard error of measurement. In addition, findings showed that MST performed better for the 1-, and 2-PL models than equivalent length CAT test. For the 3-PL model MST and CAT performed about the same.

Key words: Multi-stage tests, computerized adaptive tests, paper and pencil tests, Item Response Theory, calibration, reliability.

Introduction

Traditionally, the standard method of test delivery has been the familiar fixed length linear paper and pencil (P&P) test. However, with the advancement of technology in the past 30 years, the popularity of administering tests using adaptive methods like computerized adaptive testing (CAT) and multi-stage testing (MST) has grown in the field of measurement in both theory and practice. In CAT, items are selected for each examinee based on his or her responses to previous items in a way that targets and maximizes the precision of the estimate of the examinee's underlying latent ability.

A distinct advantage of CAT is that it offers the potential of a shorter test since items that are too easy or too difficult for an examinee are not administered, unless an item is needed to satisfy some content specification or to avoid overexposure of another item. This "tailoring" of items to an examinee's ability level leads to adaptive tests that are often more efficient than conventional P&P tests (Lord, 1980; Weiss, 1982), typically requiring examinees to answer fewer items to attain an equivalent level of precision (Green, 1983; Schnipke & Reese, 1997).

Although there are many advantages associated with CAT, there are some criticisms as well. First, examinees taking a computerized adaptive test are typically not permitted to review their answers to previous questions. Second, the number of items exposed in a computerized adaptive test is quite high (Luecht, Nungester, & Hadadi, 1996). Finally, in CAT, millions of different test forms are possible from a single item pool and it is, therefore, not feasible for people to review every test form for quality assurance purposes (Luecht & Nungester, 1998).

An alternative to CAT that eliminates some of the criticisms of CAT is multi-stage testing (MST). MST is a compromise between P&P and CAT and is, in fact, a special case of CAT that allows for item review, reduces the number of items exposed, makes the implementation of quality assurance more feasible, and still maintains all of the advantages of a test delivered via the computer.

In MST, there is partial adaptation of the test to individual examinees. Rather than adapting the test to individuals item by item as in CAT, the test adapts to examinees in stages. In MST, all examinees are administered a common set of items known as a routing or stage-one test. Depending on examinee performance, the examinee is routed to one of several alternative second-stage tests, each of which consists of a fixed set of items and differs on average difficulty. Depending upon examinee performance on the second-stage test, he or she is routed to one of several alternative third-stage tests. This process continues depending on the number of stages in the MST procedure. The number of stages and the number of blocks per stage, among other factors, vary between different testing programs that utilize MST.

While MST appears to eliminate some of the common criticisms of CAT, inherent in MST procedures are two drawbacks: the potential decrease in accuracy of ability estimation and a likely loss of efficiency relative to CAT (Kim & Plake, 1993; Luecht et al., 1996; Schnipke & Reese, 1997). While these findings are derived from studies that used items rather than sets and were based on a specific IRT model, the purpose of this study was to investigate how well adaptive procedures function when a test is completely based on item *sets* and how well these procedures perform with various IRT models.

Significance of the study

In practice, there are many standardized tests that have sections that include only set-based items. To date, there is no study in the literature that compares MST and CAT procedures to P&P testing when a test is completely set-based under various IRT models. Therefore, the question of interest was which one of the testing procedures under consideration provides more accurate ability estimates and measurement precision when the test is completely set-based for the 1-, 2- and 3-PL models?

The purpose of this study was to investigate measurement precision under various testing procedures for the 1-, 2- and 3-PL models when the test is completely set-based. In particular, the comparisons of interest were MST with CAT and MST with P&P.

Method

Using a 440-item pool (64 sets) from eight paper and pencil forms of an operational MCAT Verbal Reasoning Test, this study compared a 32-item computer adaptive test with a 33-item multi-stage test and a 55-item P&P test with a 54-item multi-stage test. Each form was calibrated using the 1-, 2-, and 3-PL models in PARSCALE using marginal maximum likelihood estimation (Muraki &Bock, 1991) and then appropriately scaled to the reference form (Form 38A). Comparisons between testing procedures were made in terms of measurement precision, bias conditioned on number right true scores, content constraints, and item exposure for the 1-, 2- and 3-PL models.

Multi-Stage Test

Although there is an infinite number of MST designs as the number of stages and number of levels per stage can vary, as shown in Figure 1 a two-stage test with three levels in the second stage was used, as that is what the item pool could support. Moderately difficult sets were used to build the first stage block and easy, moderate and difficult sets were used to build the easy, moderate, and difficult blocks, respectively, in the second stage. Note that because the item pool consisted of 64 sets, for each set, information from the items was aggregated to the set level, which resulted in one information function for each set. Furthermore, to reflect what one might do in practice for test security, two forms were assembled. Thus, there was a total of eight blocks; two blocks at stage one and six blocks at stage two (i.e. two easy, two moderate, and two difficult blocks).

Insert Figure 1 about here

Comparison between MST and CAT

To allow for a fair comparison between MST and CAT, both types of tests had similar test length and content constraints. To best match the length of the 32-item CAT, the MST had a fixed length of 33 items (16 items in routing block and 17 items in each second stage block). To create tests with no more than 33 items (or 32 for CAT) and meet all content constraints, set-based items were trimmed to a length of five or six for MST and CAT. The criteria used to trim sets were homogeneity and difficulty. Table 1 presents the content constraints for CAT and MST. The values in Table 1 indicate the number of sets of each content type that are desirable for each testing procedure. Note that the MST procedure employed did not allow for control at the item level and so no item level content constraints were included in MST as they were with the CAT.

Insert Table 1 about here

Comparison of Multi-Stage and P&P Tests

For a fair comparison between MST and P&P testing, both type of tests had similar test characteristics. A 55-item P&P test was compared to 54-item MST. For the MST, the length of the routing block was 23 items and the length of the second stage blocks was 31 items. Table 2 presents the content constraints for the P&P tests and for MST.

Insert Table 2 about here

Comparison between IRT Models for Various Testing Procedures

This study utilized a number right true score metric based on one of the eight MCAT test forms as the basis of comparing the IRT models under consideration. Table 3 provides a summary of the relationships between number right true scores and theta values based on the 1-, 2-, and 3-PL IRT scales.

Insert Table 3 about here

Block Assembly

The blocks for MST were assembled independently for each of the three IRT models. The goal for the first stage and at each level of the second stage was to create two parallel blocks so that the accuracy of the estimated ability would be the same for examinees whose tests follow the same routing path. Set-based blocks were assembled based on the following procedure. First, sets for each content type were classified as easy, moderate or difficult. Next, blocks were created by matching sets together in a way that content specifications were met. Finally, reviewers selected blocks that met content specifications and measurement properties. An analytical description of the steps taken to assemble blocks follows. Note that the following procedure was carried out independently by two reviewers. In case of disagreement in the selection of sets, reviewers discussed their choices until a common decision was reached.

1. *First Stage/Routing Block*: Three sets with the most information across a wide range of abilities from each of the three content types were chosen. This resulted in 27 (3 x 3 x 3) possible blocks that met the specifications of content constraints.
2. *Second Stage Blocks*: First, sets that provided the most information in the range $\theta > 1$, $-1 \le \theta \le 1$, and $\theta < -1$ were classified as difficult, moderate, and easy sets, respectively. Then, the three most informative sets for the easy, moderate, and difficult categories were chosen. This resulted in 27 possible blocks for each of the three levels.
3. Reviewers selected the two most informative blocks that did not overlap content-wise with each other for the first stage and for each level of the second stage.

Figures 2 to 5 present the graphs of the blocks for the easy, moderate and difficult levels for the 1-, 2-, and 3-PL model, respectively. Blocks that belong in the same level provide about the same amount of information at a given ability level. Notice that for the 2-PL model, blocks in the difficult level provide less information than blocks in the easy and moderate levels. On the other hand, for the 3-PL model, blocks in the difficult level provide much more information than blocks in the easy level. This occurred because the pool did not have enough items at each level with high discrimination values to support second stage blocks with high information at each level for the 2- and 3-PL models.

Insert Figure 2 about here

Insert Figure 3 about here

Insert Figure 4 about here

Simulation

Item Pool

Item parameters from an MCAT Verbal Reasoning item pool with 440 items (a total of 64 set-based item sets of various length) were used to assemble multi-stage, computer adaptive, and P&P tests. The 440 items came from eight P&P forms of the MCAT Verbal Reasoning test.

Simulated Examinees

Abilities of 500 simulated examinees were generated at 20 number right true scores (16 to 54 in increments of 2) for MST and CAT. This resulted in a total of 10,000 examinees for each of the 1-, 2-, and 3-PL models.

MST Simulation Procedure

The following steps were used for the MST simulation:

1. Randomly select and administer a block from the first stage.
2. Estimate examinee's ability after first stage completed using the maximum likelihood procedure for a given IRT model.
3. Based on value of estimated ability from first stage, route examinee to a level that best matches his or her estimated ability and randomly select and administer one of the parallel blocks. Specifically, if the estimated ability of an examinee from first stage was $\theta < -1$, then one of the second stage easy blocks was randomly administered; if the estimated ability of an examinee from first stage was $-1 \le \theta \le 1$, then one of the second stage moderate blocks was randomly administered; and if the estimated ability of an examinee from the first stage was $\theta > 1$, then one of the second stage difficult blocks was randomly administered.
4. Obtain final ability estimate after examinee completes second stage.

CAT Simulation Procedure

The following steps were used for the CAT simulation:

1. Select and administer a moderately difficult item.
2. For a given IRT model, estimate examinee's ability using maximum likelihood estimation.

3. Based on this estimated ability, select a new item that maximizes information and meets content, exposure and overlap constraints.
4. Continue this process until an examinee is administered 32 items

Results

Results for both comparisons, MST vs. CAT and MST vs. P&P, are presented in the following order: measurement precision, bias, content constraints, and item exposure.

Measurement Precision

Measurement precision was investigated in terms of conditional standard error of measurement (CSEM) at each generated number right true score and in terms of reliability. The reliabilities were calculated based on a weighted sum of the CSEMs using the approach recommended by Green et. al. (1984).

Table 4 and Figures 5 to7 summarize reliability and the CSEMs at each of the 20 generating number right true scores for the 1-, 2-, and 3-PL models for MST and CAT. As indicated by the greater values of reliability for the 2- and 3-PL models relative to the 1-PL model for both CAT and MST, CSEMs were smaller for the 2- and 3-PL models than the 1-PL model. This is due to the 2- and 3-PL models taking into account discrimination (and guessing for 3-PL) when selecting items, whereas the 1-PL model assumes all items are equally discriminating. In terms of comparing measurement precision between CAT and MST, the two testing procedures were similar, with MST being only slightly more precise. This was somewhat unexpected given the differing levels of adaptation between CAT and MST. CAT adapted item by item whereas MST only adapted once between stages. The difference may be due to the increased item content constraints that were placed on the CAT, but not on the MST.

Insert Table 4 about here

Insert Figure 5 about here

Insert Figure 6 about here

Insert Figure 7 about here

Table 5 and Figures 8 to 10 summarize reliability and CSEMs at each of the 20 generating number right true scores for the 1-, 2-, and 3-PL models for MST and P&P testing. Results indicate that the MST procedure resulted in similar or slightly smaller CSEMs and hence greater or equal reliability for all three models than was observed with P&P testing. This was expected given the adaptive nature of MST.

Insert Table 5 about here

Insert Figure 8 about here

Insert Figure 9 about here

Insert Figure 10 about here

Table 6 summarizes the reliability of scores of the different testing lengths and procedures. In summary, the reliability for the MST procedure was slightly better than an equal length CAT procedure for the 1- and 2-PL models and was the same for the 3-PL model (R=0.85).

As was expected, for a longer length MST (54-items), the reliability value was greater than the shorter length MST (33-items) for all three models. Note however, that for the 33-item MST procedure, the reliability value for the 2-PL model was exactly the same as the reliability value for the 3-PL model (R= 0.85 for both models). This pattern was not observed for the 54-item MST. For this length, the 3-PL model reported the highest reliability value (R=0.88), with the 2-PL model being second best.

In addition, the MST procedure reported higher reliability values than an equal length P&P testing procedure for the 2- and 3-PL models. For the 1-PL model, both testing procedures, MST and P&P, had exactly the same reliability value. Finally, for the P&P testing procedure all models reported the same reliability value (R=0.85).

Insert Table 6 about here

Bias

This section presents the results of the comparison of CAT and MST, as well as IRT models with respect to the bias values at the 20 generating number right scores. Note that the bias values were not computed for the P&P test since the results of this procedure were not simulated. Therefore the only comparison presented is CAT vs. MST.

As shown in Figures 11 and 12, neither the CAT nor MST procedures overly biased scores or had any systematic bias.

Insert Figure 11 about here

Insert Figure 12 about here

Figures 13 to 15 present the bias values at the 20 generating number right scores for the CAT and MST procedures for the 1-, 2-, 3-PL models, respectively. For no model was one testing procedure better than the other.

Insert Figure 13 about here

Insert Figure 14 about here

Insert Figure 15 about here

Content Constraints

Sometimes CAT, in its attempt to maximize reliability and satisfy exposure specifications simultaneously, violates content constraints. For MST, the way blocks are assembled does not allow for any content violations. However, in MST, it is difficult to incorporate item level content constraints and so they were not incorporated, yet they were in CAT and they were violated as delineated in Table 7. A fairer comparison of CAT and MST with regard to content violations would be to rerun the CAT simulation with only set constraints.

Insert Table 7 about here

Item Exposure

Item exposure rates are reported based on the134-item pool for the MST procedure and the 440-item pool for the CAT procedure. The item exposure rate for the MST procedure followed a different pattern than the CAT procedure. Figures 16 and 17 summarize item exposure rates for the CAT and MST procedures, respectively. Each figure presents the relationship between the cumulative percent of the items in the pool and the item exposure rate. This relationship was the same for all IRT models under consideration from both testing procedures. From Figure 17, it can be observed that up to 51% (68 of 134 items) of the items in the pool had exposure rate of less than 0.1 and about 24% (32 of 134 items) of the items in the pool had exposure rate as high as 0.50. On the other hand, using the CAT procedure, up to 70% (325 of 440 items) of the items in the pool had exposure rates less than 0.1 and no item in the pool had exposure rate higher than 0.3.

In summary, using the CAT procedure there are more items exposed at low exposure rates and there were no items in the pool that had exposure rate as high as 0.50.

Insert Figure 16 about here

Insert Figure 17 about here

Conclusion

This study compared equal length set-based computerized adaptive tests and multi-stage tests, as well as equal length set-based multi-stage tests and P&P tests. From the results of this study, the MST procedure performed better than an equal length P&P testing procedure with respect to measurement precision. In addition, the 32-item CAT and 33-item MST provided the same reliability as a 55-item P&P test. Furthermore, the 33-item MST possessed slightly higher reliability than the 32-item CAT test for the 1- and 2-PL models, while both testing procedures had exactly the same reliability for the 3-PL model.

In addition, there were no content violations at the set-based level for any of the testing procedures. Furthermore, the CAT procedure reported more items with lower exposure rate than the MST testing which reported fewer items for the same low exposure rates but some of the items in the MST pool had higher exposure rate than any of the items in the CAT pool. Also, the number of items in a pool required to deliver MST is much smaller than the number of items to deliver CAT.

Keeping in mind the above findings, the recommended testing procedure for a set-based test is the 33-item MST procedure. In addition, based on the reliability value as well as the findings from the bias analysis, the recommended model for the 33-item MST is the 2-PL model. However, the reason to choose the 2-PL model rather than the 3-PL model should not only be determined by the results of the measurement precision, but also whether the guessing parameter provides some important and meaningful information about the test. In addition, if the cost of having more items does not play as an important factor as the precision of the test, then the recommended testing procedure is the 54-item MST test. For this length of MST, the recommended model is the 3-PL which provides the highest reliability value. Finally, if the high exposure of some of the items in the MST procedure is of a concern then the use of more blocks in the first stage is one way to reduce the high exposure rates of these items.

References

Green, B. F. (1983). The promise of tailored tests. In H. W. Wainer and S. Messick (Eds.), Principals of modern psychological measurement (pp. 69-80). Hillsdale, NJ: Lawrence Erlbaum Associates.

Kim, H., & Plake, B. S. (1993, April). Monte Carlo simulation comparison of two-stage testing and computerized adaptive testing. Paper presented at the meeting of the National Council on Measurement in Education, Atlanta, GA.

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

Luecht, R. M. Nungester (1998). Some practical examples of computer adaptive sequential testing. Journal of Educational Measurement, 35, 229-249.

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

Olsen, J. B., Maynes, D. M., & Slawson, D. A., (1986, April). Comparison and equating of paper-administered, computer-administered and computerized adaptive tests of achievement. Paper presented at the meetings of the American Educational Research Association, San Francisco, CA.

Schnipke, D. L., & Reese, L. M. (1997, March). A comparison of testlet-based test designs for computerized adaptive testing. Paper presented at the meeting of American Educational Research Association, Chicago, IL.

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

Table 1

Content Constraint for CAT and MST

		MST		
Constraint	CAT	First Stage	Second Stage	Total MST
S:Human	2	1	1	2
S:NatSci	2	1	1	2
S:SocSci	2	1	1	2
S:Six	2	1	2	3
S:Five	4	2	1	3
I:Comp	8-12	-	-	-
I:Eval	4-8	-	-	-
I:Appl	7-10	-	-	-
I:Incorp	6-9	-	-	-
I:Human	10-12	-	-	-
I:NatSci	10-11	-	-	-
I:SocSci	10-12	-	-	-

Table 2

Content Constraints for P&P and MST

		MST		
		First	Second	Total
Constraint	P&P	Stage	Stage	MST
S:Human	2-4	1	2	3
S:NatSci	2-2	1	1	2
S:SocSci	2-3	1	2	3
S:Ten	1-1	0	0	0
S:Eight	0-1	2	0	2
S:Seven	1-3	1	1	2
S:Six	3-5	0	4	4

Table 3

Theta-to-Number Right True Score for the 1-, 2-, and 3-PL Models

NR True	1-PL θ	2-PL θ	3-PL θ
54	3.5545	6.6723	4.4161
52	2.3832	3.8516	2.7141
50	1.7992	2.6623	2.0336
48	1.3891	1.9132	1.5718
46	1.0633	1.3695	1.2050
44	0.7871	0.9420	0.8942
42	0.5431	0.5870	0.6199
40	0.3212	0.2798	0.3694
38	0.1153	0.0055	0.1333
36	-0.0793	-0.2458	-0.0960
34	-0.2656	-0.4806	-0.3248
32	-0.4461	-0.7039	-0.5592
30	-0.6229	-0.9192	-0.8053
28	-0.7978	-1.1296	-1.0705
26	-0.9725	-1.3376	-1.3641
24	-1.1487	-1.5459	-1.6993
22	-1.3280	-1.7571	-2.0954
20	-1.5124	-1.9740	-2.5869
18	-1.7043	-2.2003	-3.2537
16	-1.9063	-2.4403	-4.3797

Table 4

CSEMs for 32-Item CAT and 33-Item MST for 1-, 2-, and 3-PL Models

True Score	1-PL CAT	1-PL MST	2-PL CAT	2-PL MST	3-PL CAT	3-PL MST
54	1.02	0.99	1.09	1.08	0.88	1.02
52	1.72	1.78	1.54	1.61	1.51	2.14
50	2.23	2.18	1.82	1.98	1.79	2.56
48	2.75	2.49	2.23	2.21	2.10	2.43
46	3.00	2.83	2.61	2.58	2.29	2.28
44	3.19	3.38	2.90	2.71	2.52	2.54
42	3.43	3.34	3.06	2.85	2.58	2.77
40	3.85	3.47	3.28	2.95	3.27	2.78
38	3.72	4.10	3.31	3.30	3.12	3.19
36	4.10	4.02	3.67	3.25	3.49	3.26
34	4.08	4.11	3.38	3.57	3.44	3.37
32	4.25	4.09	3.73	3.30	3.44	3.12
30	4.24	4.13	3.54	3.11	3.24	2.95
28	4.36	3.91	3.33	3.13	3.30	3.14
26	4.27	4.15	3.43	3.33	3.44	3.24
24	3.94	4.01	3.12	3.21	3.19	3.20
22	4.25	4.37	3.07	3.29	3.03	3.48
20	4.07	4.15	3.10	3.09	3.10	3.26
18	4.00	3.91	3.12	2.76	2.52	2.87
16	3.95	3.79	3.25	2.77	1.92	2.64
Reliability	0.78	0.79	0.84	0.85	0.85	0.85

Table 5

CSEMs for 54-item MST and 55-item P&P for 1-, 2-, and 3-PL Models

True Score	1-PL P&P	1-PL MST	2-PL P&P	2-PL MST	3-PL P&P	3-PL MST
54	0.98	0.78	0.86	0.99	0.82	0.80
52	1.64	1.43	1.46	1.23	1.45	1.23
50	2.05	1.83	1.82	1.61	1.84	1.61
48	2.36	1.89	2.10	1.84	2.13	1.67
46	2.59	2.37	2.34	2.10	2.35	1.93
44	2.78	2.54	2.54	2.31	2.53	2.04
42	2.93	2.91	2.70	2.67	2.67	2.47
40	3.06	3.02	2.84	2.72	2.80	2.64
38	3.16	3.26	2.96	2.88	2.90	2.75
36	3.24	3.21	3.05	3.12	2.99	2.82
34	3.30	3.30	3.12	3.08	3.06	2.87
32	3.35	3.32	3.17	3.11	3.11	2.81
30	3.37	3.41	3.21	2.96	3.14	2.98
28	3.38	3.31	3.22	3.15	3.15	2.79
26	3.37	3.40	3.22	3.07	3.13	2.90
24	3.35	3.12	3.21	3.06	3.07	2.73
22	3.32	3.31	3.18	2.93	2.94	2.82
20	3.26	3.25	3.13	3.00	2.74	2.57
18	3.19	3.17	3.06	2.81	2.50	2.27
16	3.10	2.90	2.97	2.73	2.14	1.59
Reliability	0.85	0.85	0.85	0.87	0.85	0.88

Table 6

Reliability for 1-, 2-, and 3-PL Models for Various Testing Procedures

Testing Procedure	1-PL	2-PL	3-PL
32-item CAT	0.78	0.84	0.85
33-item MST	0.79	0.85	0.85
54-item MST	0.85	0.87	0.88
55-item P&P	0.85	0.85	0.85

Table 7

Content Constraint Violations for CAT

Content	Targeted #Items			%Violations			Min. Adm.			Max. Adm		
Constraint	Low	High	Wght.	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL	1-PL	2-PL	3-PL
S:Human	2	2	10	0	0	0	2	2	2	2	2	2
S:NatSci	2	2	10	0	0	0	2	2	2	2	2	2
S:SocSci	2	2	10	0	0	0	2	2	2	2	2	2
S:Six	2	2	10	0	0	0	2	2	2	2	2	2
S:Five	4	4	10	0	0	0	4	4	4	4	4	4
I:Comp	8	12	90	.08	.05	.08	6	5	5	16	14	16
I:Eval	4	8	90	.02	.02	.03	3	3	3	10	9	10
I:Appl	7	10	90	.12	.08	.09	5	5	4	13	12	12
I:Incorp	6	9	90	.08	.12	.12	3	4	4	12	12	12
I:Human	10	12	10	0	0	0	10	10	10	12	12	12
I:NatSci	10	11	10	.11	.34	.38	10	10	10	12	12	12
I:SocSci	10	12	10	0	0	0	10	10	10	12	12	12

Figure 1

MST Design

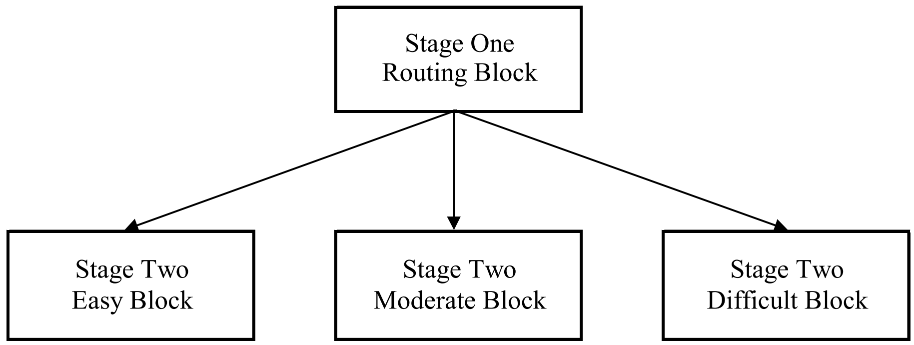

Figure 2

*Easy, Moderate and Difficult Blocks for the **1-PL** Model*

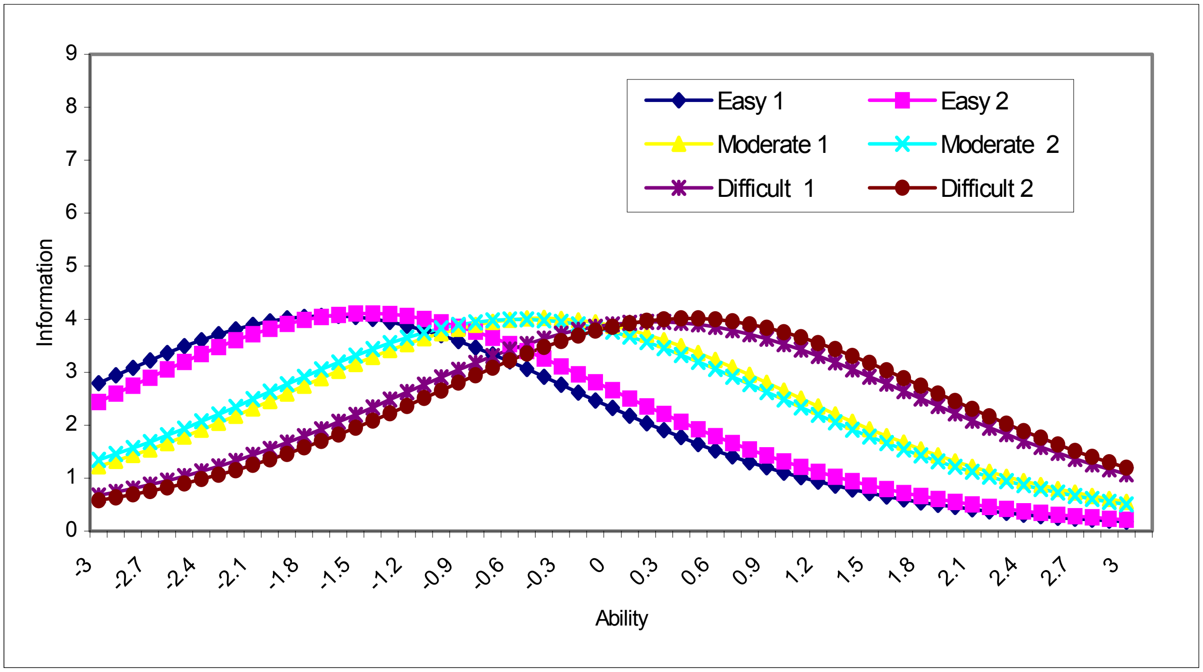

Figure 3

Easy, Moderate and Difficult Blocks for the ***2-PL*** *Model*

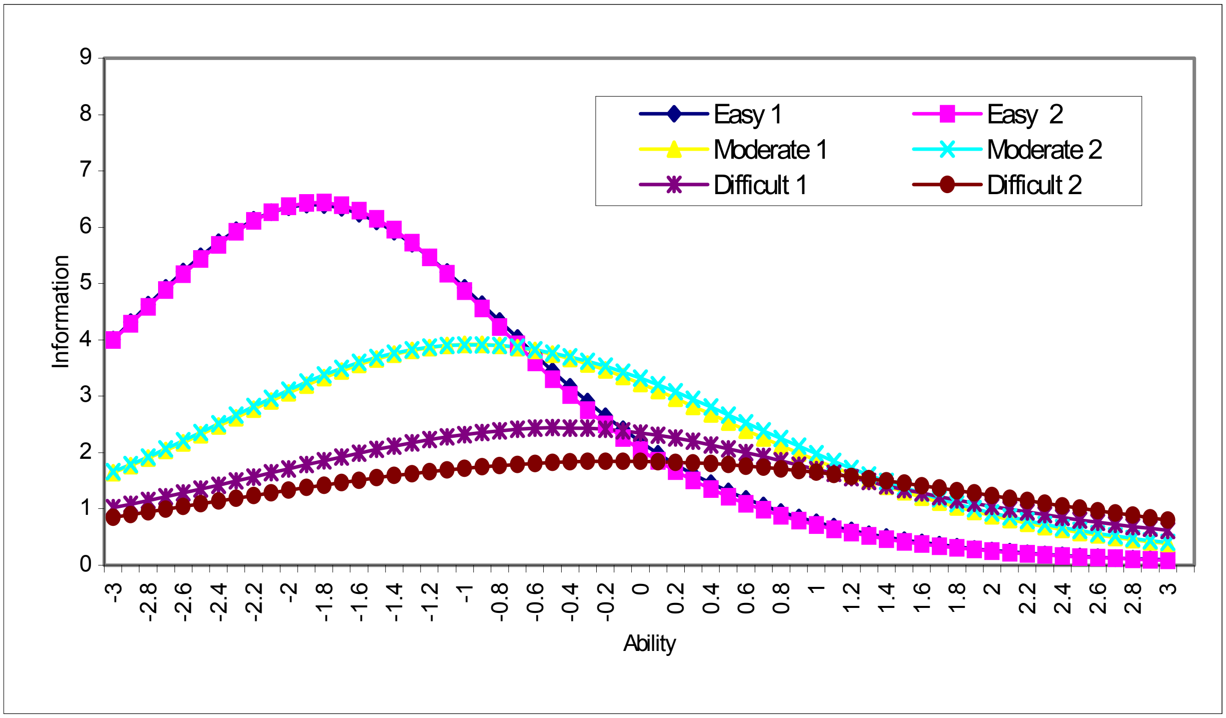

Figure 4

*Easy, Moderate and Difficult Blocks for the **3-PL** Model*

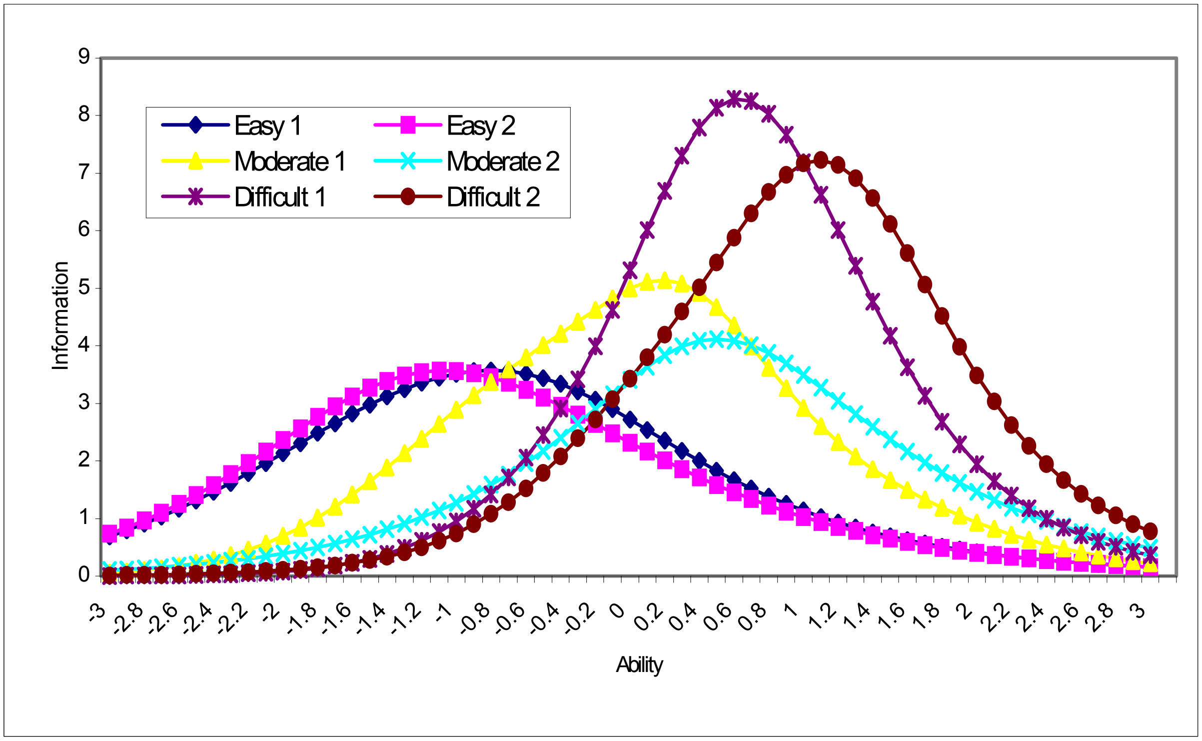

Figure 5

*CSEMs for 32-item CAT and 33-item MST for **1-PL** Model*

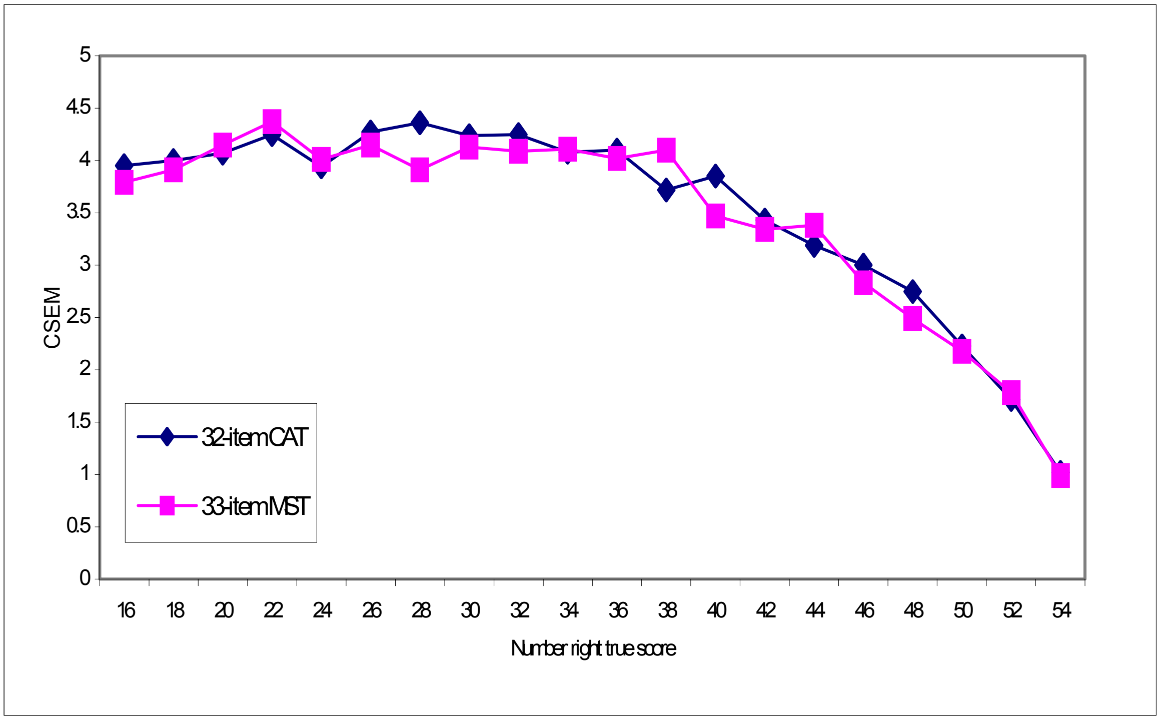

Figure 6

*CSEMs for 32-item CAT and 33-item MST for **2-PL** Model*

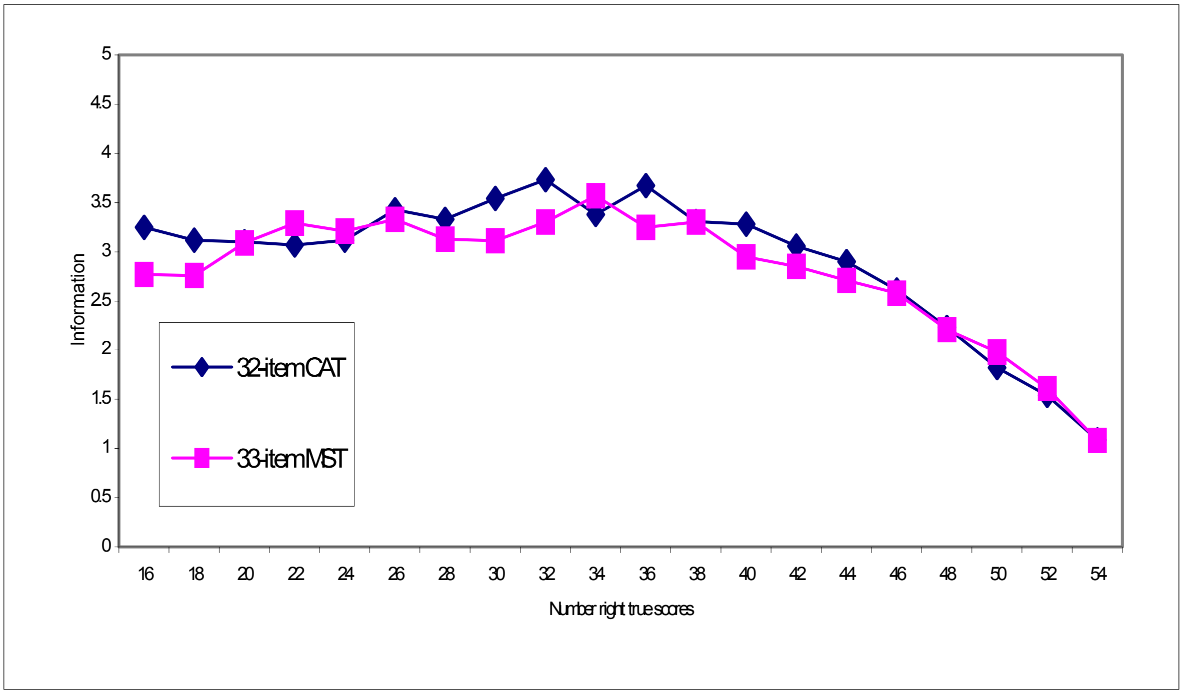

Figure 7

*CSEMs for 32-item CAT and 33-item MST for **3-PL** Model*

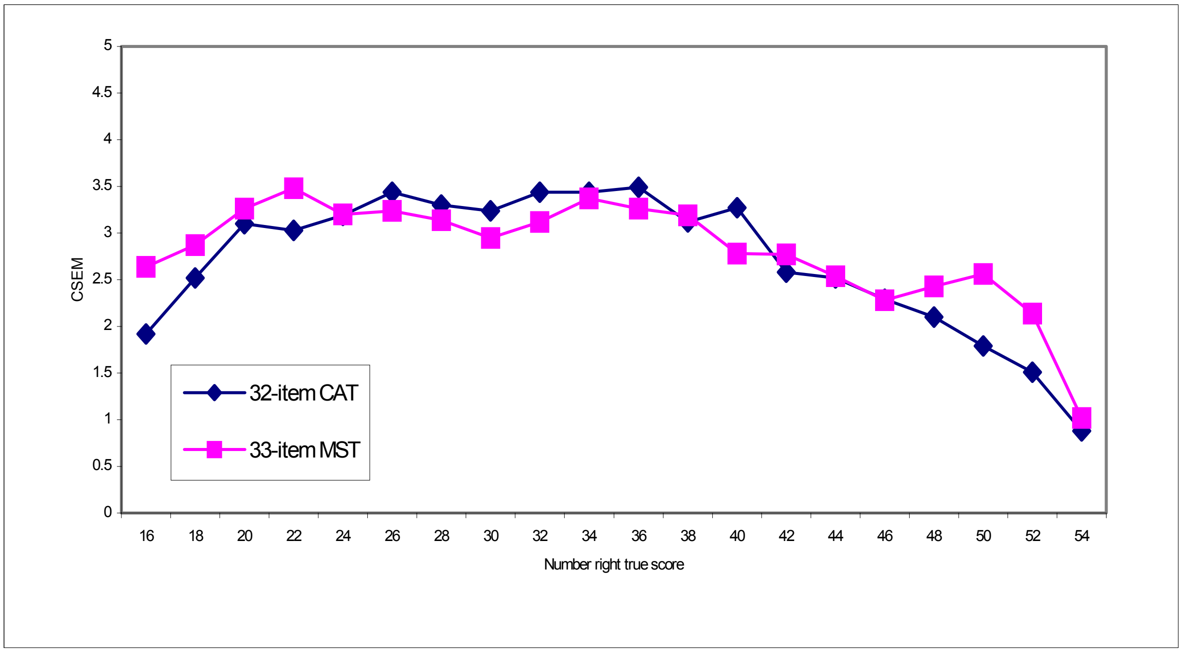

Figure 8

*CSEMs for 55-item P&P and 54-item MST for **1-PL** Model*

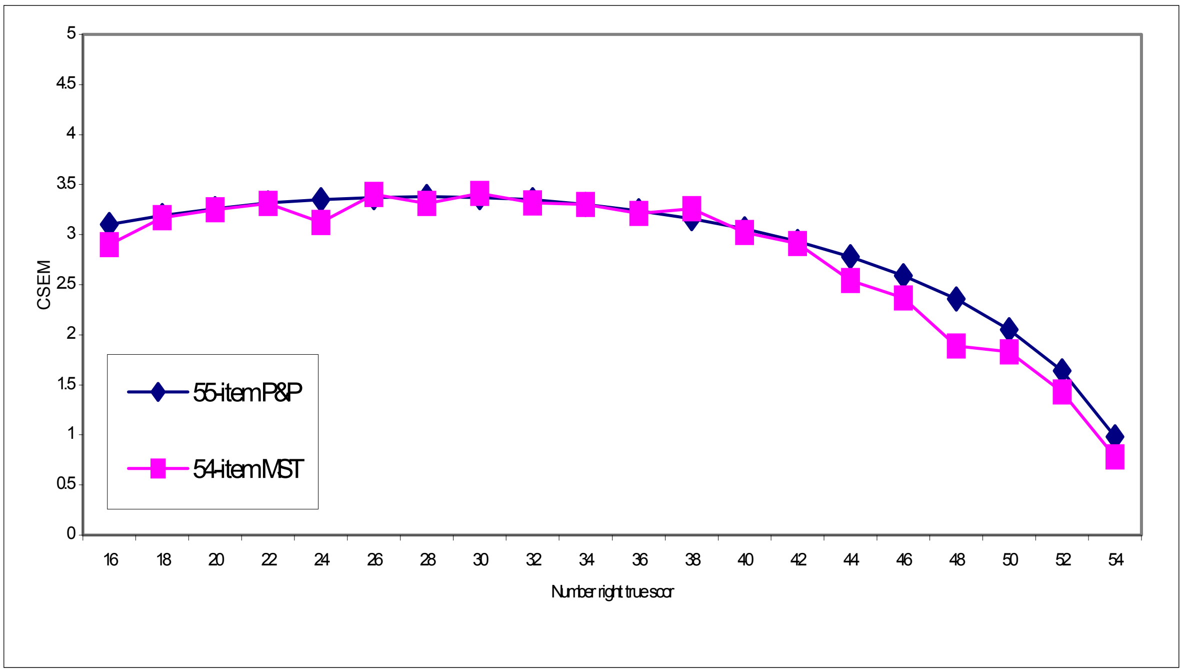

Figure 9

CSEMs for 55-item P&P and 54-item MST for ***2-PL*** *Model*

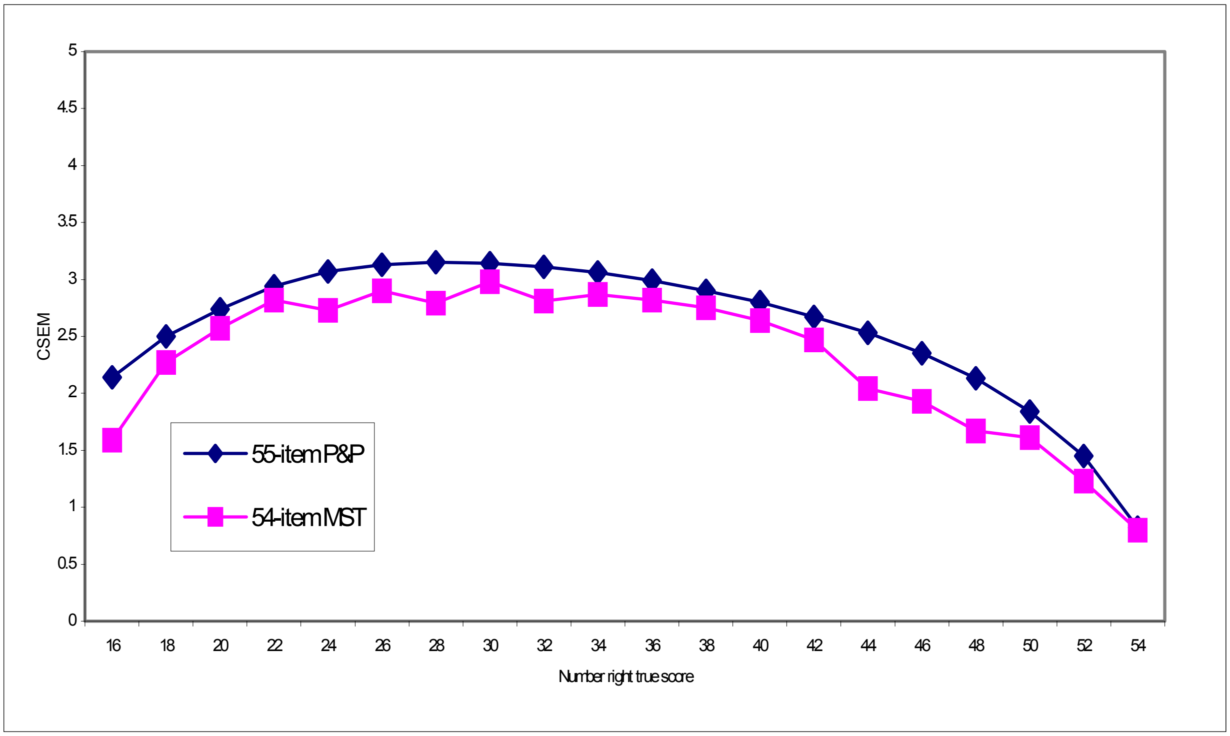

Figure 10

*CSEMs for 55-item P&P and 54-item MST for **3-PL** Model*

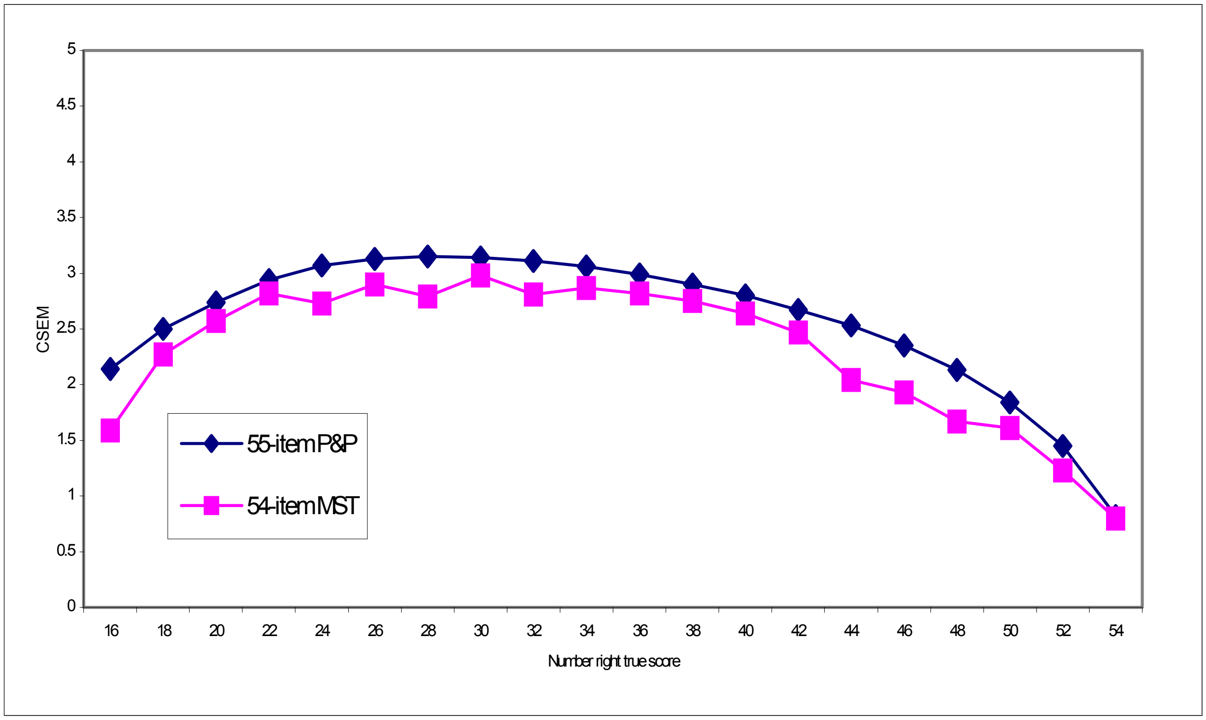

Figure 11

Bias for 1-, 2-, and 3-PL Models for 32-item CAT

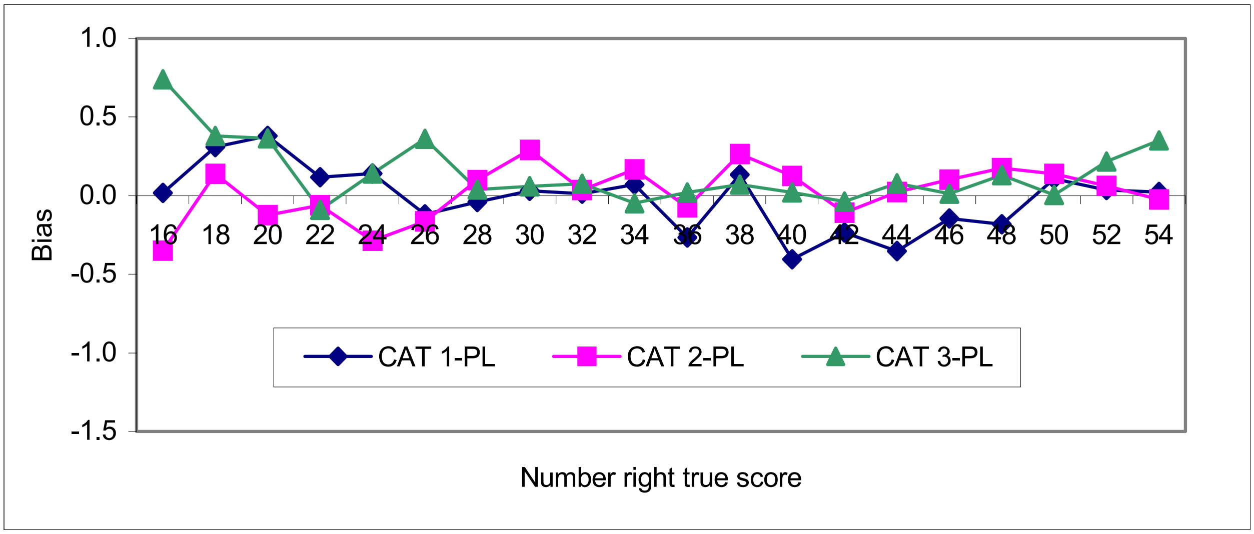

Figure 12

Bias for 1-, 2-, and 3-PL Models for 33-item MST

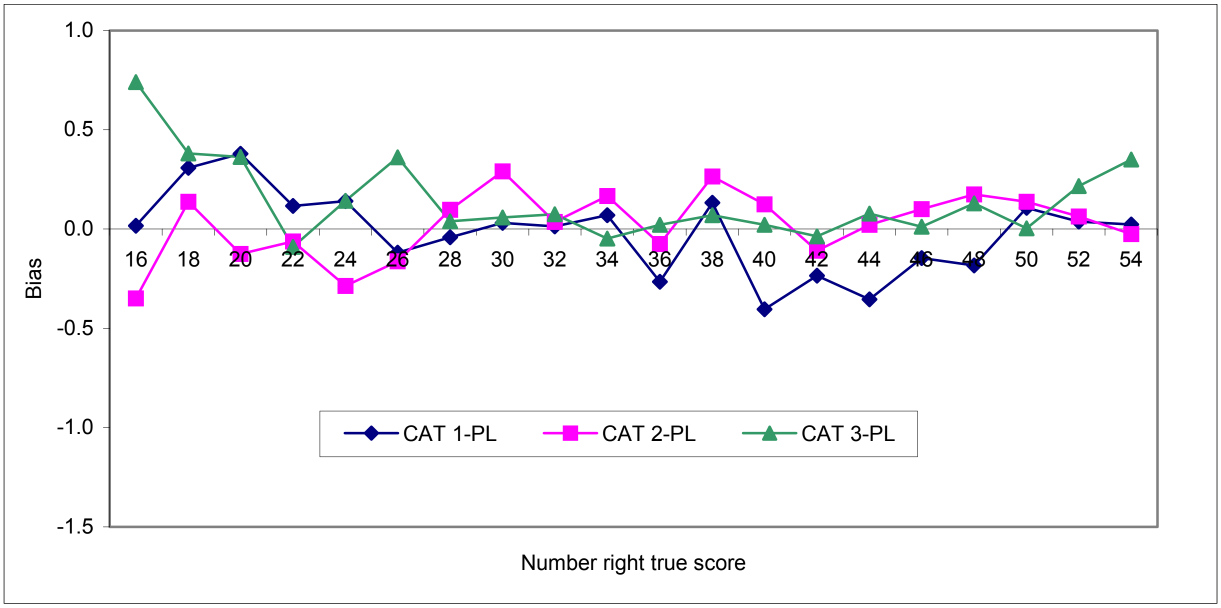

Figure 13

Bias for CAT and MST for 1-PL Model

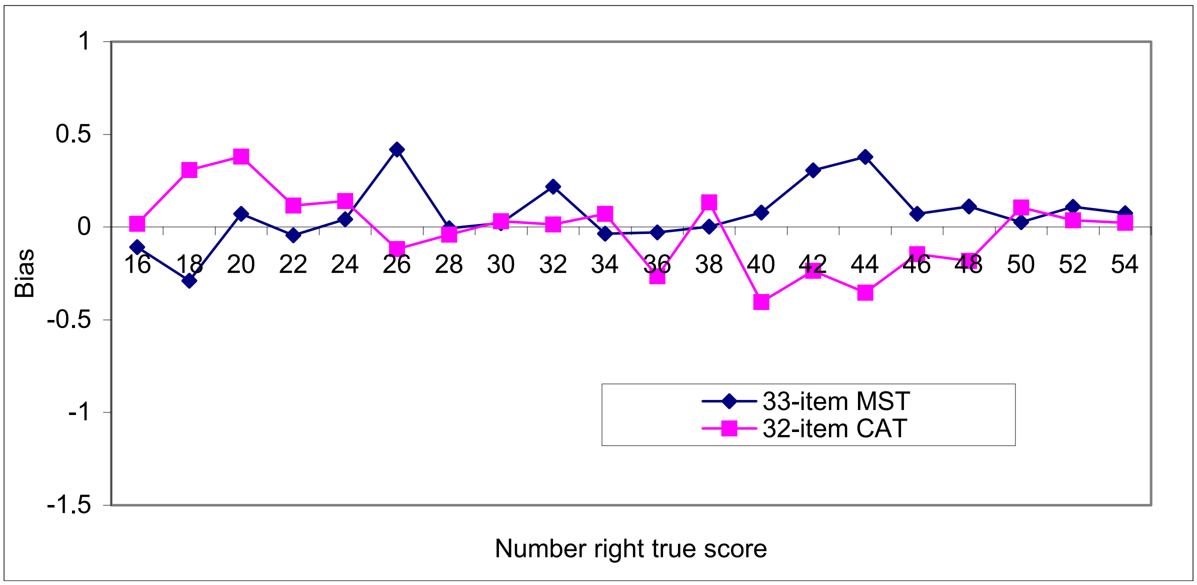

Figure 14

Bias for CAT and MST for 2-PL Model

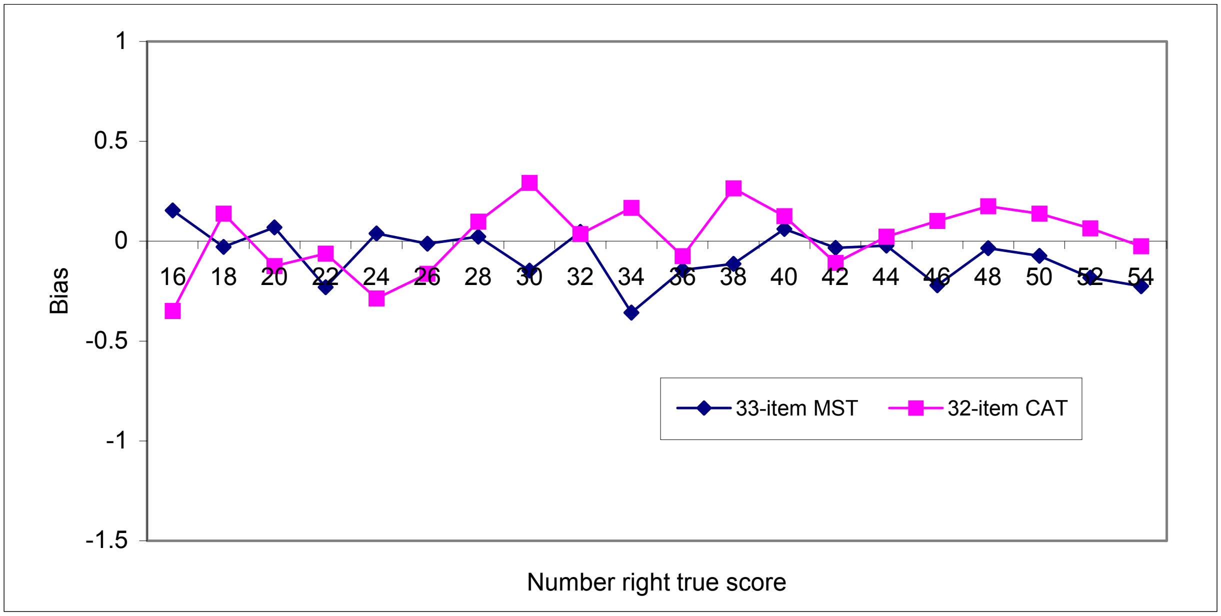

Figure 15

Bias for CAT and MST for 3-PL Model

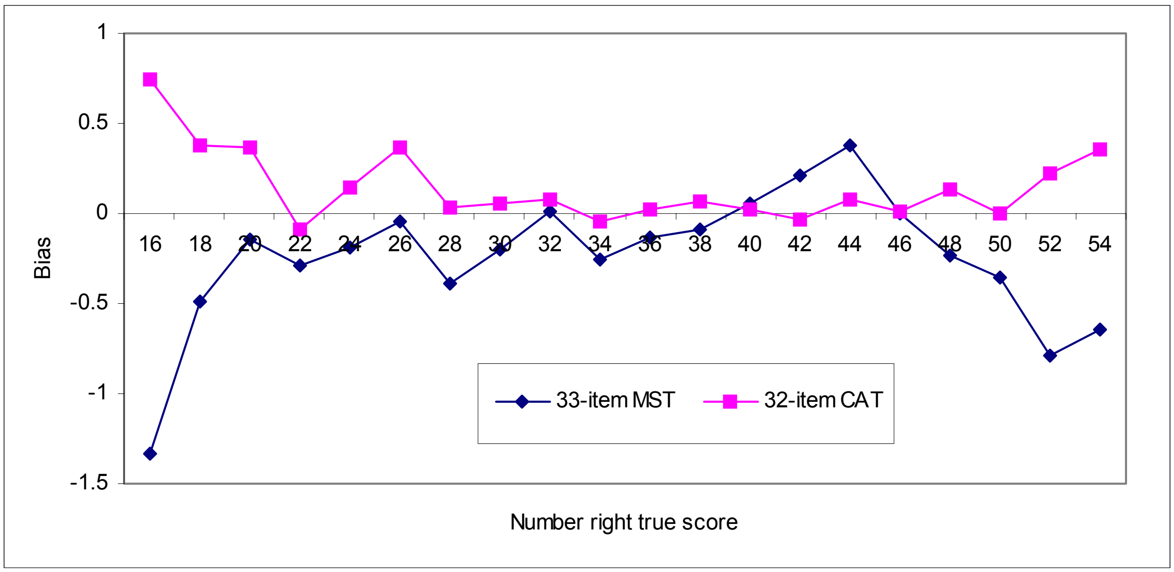

Figure 16

Item exposure rates for CAT procedure.

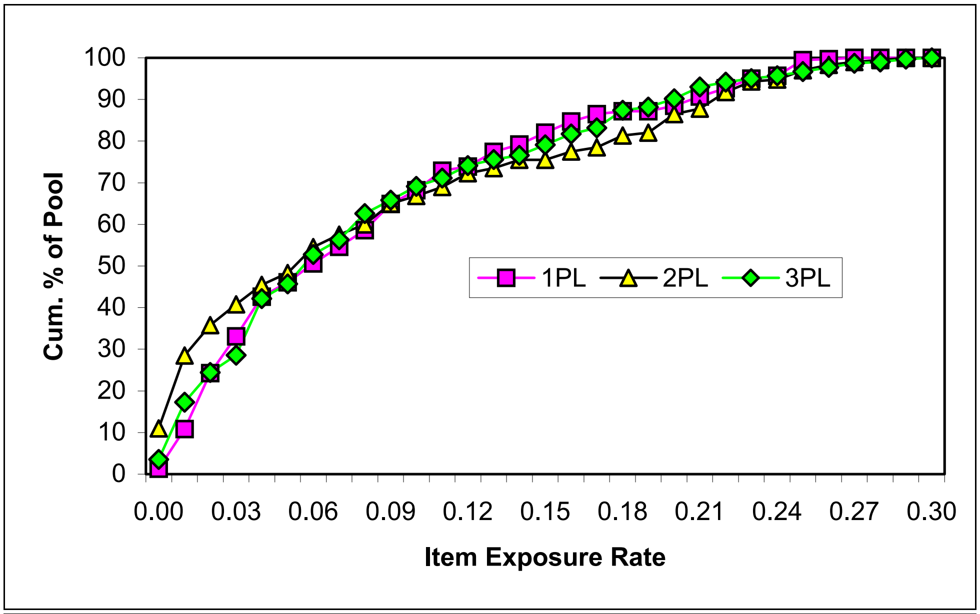

Figure 17

Item Exposure Rate for the MST procedure.