Evaluating Computer Adaptive Testing Design for the MCAT with Realistic Simulated Data[1]

by

Ying Lu, Mary Pitoniak
University of Massachusetts Amherst

Saba Rizavi, Walter D. Way & Manfred Steffen[2]
Educational Testing Service

[1] Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

[2] The authors are thankful to Drs. Tim Davey and Liane Patsula of Educational Testing Service for their valuable advice throughout this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. Introduction

The transition from paper-based testing to computer-adaptive testing (CAT) requires evaluation of different types of designs for delivery of test items. Specifically, a CAT evaluation framework should include an ideal way of depicting the real administration of a CAT and a model that is most suitable for the particular examinee population for whom CAT is intended. The purpose of this study was to evaluate CAT designs for use with the Verbal Reasoning (VRS) measure of the Medical College Admissions Test (MCAT) using realistically simulated data. In addition to model-based data generation commonly used in simulations, this study utilized more realistic simulations based on response data generated by using simulees' probabilities of correct response on conditional *observed* proportions correct.

In a typical simulation context in which model-based data generation is used, it is difficult to compare alternative IRT models as usual evaluation criteria involve comparisons between true and estimated parameters associated with the given model of interest. Empirically-based simulations help overcome this problem by providing a less model-dependent basis for model comparisons as data are generated from observed probabilities obtained on the actual data, which, although still affected by model-based parameter calibrations, are based upon realistic response patterns and should be similar across models.

One-, two-, and three- parameter model based CATs were evaluated that were designed to be parallel to the fully set based paper and pencil (P&P) MCAT Verbal Reasoning test. These results were compared using the two simulation procedures. The empirical- and model-based simulations for the three models were carried out for the situations where sets were a) trimmed to increase within-set homogeneity and b) not trimmed, so had heterogeneous difficulty distributions. Common specifications for the designs included item selection methodology using the weighted deviations model (WDM; Stocking & Swanson, 1993) and multinomial exposure control (Stocking & Lewis, 2000).

2. Significance of the Study

A limitation of adaptive testing simulations is that the data are generated using the same IRT model that is used for scoring. As mentioned before, since real data seldom fit assumed IRT

models perfectly, model-based simulations may provide incomplete and possibly misleading information about a particular adaptive testing design.

As part of the item calibrations in this study, data to help assess model-data fit were generated for each item. The item calibration procedures in PARSCALE utilized the Bock and Aitkin (1981) marginal maximum likelihood algorithm to estimate item parameters. This estimation procedure resulted in an estimated posterior distribution of ability, which was represented by discrete ability levels (quadrature points) and proportions (quadrature weights). The fit of the model for a particular item j was evaluated over the q quadrature points, where n_{jk} is the estimated number of examinees at ability level X_k and r_{jk} is the expected number of correct responses at ability level X_k. These values were used to compute chi-square fit statistics and graphical displays of model-data fit for each item. Figure 2.1 displays an example of the graphical model-data fit for an item based on a 1-PL calibration. The continuous line is the model-based item characteristic curve and the boxes represent r_{jk}/n_{jk} at each quadrature point.

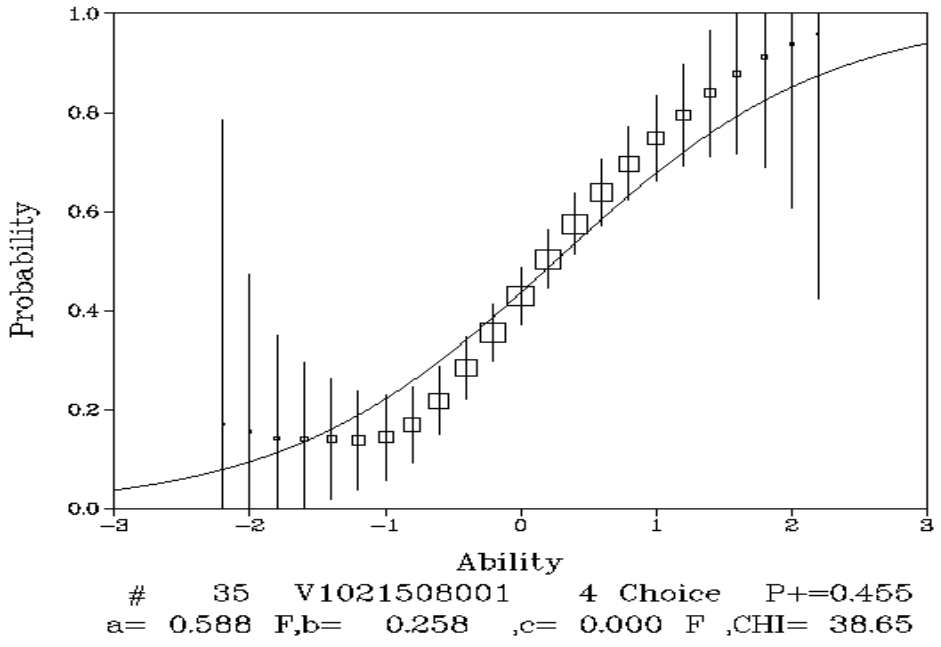

Figure 2.1: Example of a graphical model-data fit plot.

A substantial degree of model misfit is exhibited in Figure 2.1, which is not unusual in applications of the 1-PL model. In this case, the empirical data suggests a more discriminating

item than would be predicted by the 1-PL model. It appears that the performance of low ability examinees is over-predicted by the model and the performance of high ability examinees is under-predicted.

In an adaptive testing situation, the potential impact of the model-data misfit illustrated in Figure 2.1 is unclear. Ideally, this item would be administered adaptively to examinees with abilities in the range of 0.25, which is the region of the ability scale where the fit of the model to the data is good. In the *ideal* situation, the misfit at the extremes of the scale would be of no consequence. However, factors such as item pool depth, content constraints, and exposure could result in repeated adaptive administrations of this item at more extreme ability levels, in which case the model-data misfit might contribute to bias in examinees' final ability estimates who received this item. This problem is exacerbated in CAT because not all examinees receive the same items and so not all people are affected equally. Across an entire adaptive testing item pool, the patterns of model-data misfit would interact with the item selection algorithm in complex ways that are difficult to predict. By conducting the empirically-based adaptive testing simulations, one can see the impact of model-data misfit.

3. Method

Simulations were carried out using 1-, 2-, and 3-parameter models using model-based and empirically-based data generation. The empirical- and model-based simulations for the three models were carried out for the situations where sets were a) trimmed to increase within-set homogeneity and b) not trimmed so sets had heterogeneous difficulty distributions. Our interest in the simulation results fell into four categories: measurement precision of the CATs; overall reliability of the CATs; exposure rates of set stimuli (passages) and items within sets; and how well the content specifications for the simulated CATs were satisfied.

3.1 Item Pool

The study utilized an item pool consisting of items from eight paper and pencil forms of the MCAT Verbal Reasoning test. Each form was calibrated using the one-, two-, and three-parameter logistic model using PARSCALE (Muraki & Bock, 1991). Simulations using

different IRT models were conditioned on the same number right-true score metric based on one of the eight test forms so that simulation results could be compared across models.

Item pools were first created by deleting items that exhibited poor model-data fit and items that differed in difficulties from the other items within a set. The decision of set trimming was made to make the sets more homogeneous in difficulty, and to alleviate the adverse impact of using the set as administration unit in CAT on measurement precision. However, the disadvantage of set trimming is that the satisfaction of content specification and item exposure rate requirement might be negatively affected. Therefore the simulations were repeated with sets having more heterogeneously difficult distributions for within-set items, that is, without trimming the sets. These decisions were made independently for each IRT model. The number of items recovered from the 1-, 2-, and 3-PL item pools were 35, 38, and 40, respectively for the non-trimmed scenario. It was expected that not trimming would adversely affect measurement precision, however, it might increase the proportion of cases for which the desired content specifications were satisfied.

3.2 Content Specifications

A fixed length of 32 questions (four 5-item passages and two 6-item passages) was determined for the adaptive test after analyzing the content specifications for the 55-item paper-and-pencil VRS exam. Because of the set-based nature of the measure, it was impossible to reduce the number of items and passages within content areas proportionally to the reduction in total items administered. Table 3.2.1 lists the content specifications used in the simulations. There were five content constraints for stimuli and seven content constraints for items. The goal for each adaptive test was to have the number of items associated with each content constraint range from the lower bound to the upper bound. The last column indicates the weight assigned to meeting each constraint. The item-level cognitive constraints carried the highest weights because they were the most difficult constraints to meet. Three additional constraints (not included in the table), called “poor fit”, “medium fit”, and “good fit”, were defined for the pools based on each model. The number of items in the lower and upper bounds differed by model. These constraints attempted to control the inclusion of poor fitting items in the adaptive tests, although this goal was not considered as important as satisfying the content constraints.

Table 3.2.1: Verbal Reasoning Adaptive Test Content Constraints

Content Constraint	Lower Bound	Upper Bound	Weight
S:Human	2	2	10
S:NatSci	2	2	10
S:SocSci	2	2	10
S:Six	2	2	10
S:Five	4	4	10
I:Comp	8	12	90
I:Eval	4	8	90
I:Appl	7	10	90
I:Incorp	6	9	90
I:Human	10	12	10
I:NatSci	10	11	10
I:SocSci	10	12	10

In addition, item information at the estimated ability level at each point in the test was also considered a constraint. Specifically, the algorithm attempted to maximize information subject to the other (content) constraints. The conditional exposure rate for an item (conditioned on ability level) was targeted at a maximum of 0.25.

3.3 Simulated Examinees

Data for 500 simulees were generated at 20 number right true scores (from 16 to 54 in the increments of 2) using model-based and empirically-based data generation.

3.3.1 Model-Based Data Generation

In typical adaptive testing simulations, item and ability parameters for a particular IRT model are used to generate data according to well-known data generation procedures. In these procedures, an initial ability is first assumed, and the following steps occur in an iterative fashion:

1. The adaptive item selection algorithm chooses an item appropriate for the initial assumed ability level.
2. Using the true item and ability parameters and the chosen IRT model, the probability of a correct response, $P(\theta)$, is generated.

3. P(θ) is compared to a random uniform deviate between 0.0 and 1.0.
4. If P(θ) is greater than or equal to the random deviate, a correct response is generated; otherwise, an incorrect response is generated.
5. An updated ability estimate is obtained using the responses generated so far and the item parameters comprising the pool (usually these are the same as the true item parameters).
6. The process is repeated until an appropriate stopping rule is satisfied (e.g., a fixed number of items are administered or a fixed level of precision is achieved).

3.3.2 Empirically-Based Data Generation

In case of empirically-based data generation, examinee responses were generated not according to the model-based probability of a correct response, but according to an observed conditional probability of a correct response obtained from the actual data, which we call the empirically-based probability. The following adaptive testing procedures were carried out using the data generation based on empirical probability:

1. The adaptive selection algorithm selected an item appropriate for the initial assumed ability level.
2. Using the true ability parameters and linear interpolation, an empirically based P(θ) was generated by entering a table that consists of the r_{jk}/n_{jk} at each ability level (quadrature point) from the original item calibrations using PARSCALE.
3. P(θ) was compared to a random uniform deviate between 0.0 and 1.0.
4. If P(θ) was greater than or equal to the random deviate, a correct response was generated; otherwise, an incorrect response was generated.
5. An updated ability estimate was obtained using the responses generated so far and the item parameters comprising the pool (in this case, the calibrated item parameters).
6. The process was repeated for a fixed test length.

As small values of n_{jk} and r_{jk} can lead to distorted probability due to rounding errors, quadrature points were only regarded as valid if n_{jk} was larger than or equal to 0.0010, and r_{jk} was larger than or equal to 0.0001. An interval of quadrature points that were valid was identified as $[X_m, X_n]$ for every item. And the following rules were applied when generating the probability.

- When the requested ability level was within the interval of $[X_m, X_n]$, the probability was obtained using linear interpolation from the expect proportions correct at the two quadrature points that were closest to the requested ability level.

- When the requested ability level was lower than the lowest valid quadrature point, i.e. X_m, for 1-PL and 2-PL, probability was set to r_{jm}/n_{jm} at X_m if r_{jm}/n_{jm} was smaller than 0.25, and to 0.25 if otherwise, for 3-PL, the probability was set to r_{jm}/n_{jm} at X_m if r_{jm}/n_{jm} was smaller than *c* parameter, and to *c* parameter if otherwise.

- When the requested ability level was between the highest valid quadrature point, i.e. X_n, and the highest quadrature point, and when the highest quadrature point was not valid, probability was obtained using linear interpolation between r_{jn}/n_{jn} at X_n and 1.

- When the requested ability level was higher than the quadrature point range available, probability was set to 1.

- Polynomial regression was then conducted to get a smoothed function of r_{jk}/n_{jk} over X_k, so as to eliminate noise in the data due to small sample size and to avoid the imprecision introduced by linear interpolation especially at extreme ends of the posterior ability distribution (Davey, 2002).

3.4 Comparison between IRT Models and Various Data Types

Because the scales of different IRT models are unique, it is often difficult to compare across simulations based on different IRT models. In addition, generating an empirically-based P(θ) makes the usual comparisons between estimated and true parameters more complicated. These difficulties were resolved by conditioning simulations on the number right true score metric based on one of the eight MCAT test forms. The number right metric provides a basis for comparing the 1-, 2-, and 3-PL IRT scales under study, as well as interpreting results of simulations where P(θ) is generated empirically.

Table 3.4.1 summarizes the theta-to-number right true score relationships based on the 1-, 2-, and 3-PL IRT scales for 20 number right true scores on Form 38A, which served as the reference form for the simulations.

Table 3.4.1: 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	NR True	1-PL θ	2-PL θ	3-PL θ
54	3.5545	6.6723	4.4161	34	-0.2656	-0.4806	-0.3248
52	2.3832	3.8516	2.7141	32	-0.4461	-0.7039	-0.5592
50	1.7992	2.6623	2.0336	30	-0.6229	-0.9192	-0.8053
48	1.3891	1.9132	1.5718	28	-0.7978	-1.1296	-1.0705
46	1.0633	1.3695	1.2050	26	-0.9725	-1.3376	-1.3641
44	0.7871	0.9420	0.8942	24	-1.1487	-1.5459	-1.6993
42	0.5431	0.5870	0.6199	22	-1.3280	-1.7571	-2.0954
40	0.3212	0.2798	0.3694	20	-1.5124	-1.9740	-2.5869
38	0.1153	0.0055	0.1333	18	-1.7043	-2.2003	-3.2537
36	-0.0793	-0.2458	-0.0960	16	-1.9063	-2.4403	-4.3797

4. Results

The following sections present the results of the study in the following sequence: a) model-based simulations with set trimming for the three models; b) model-based simulations without set trimming for the three models along with the comparisons with the previous

simulations with set trimming; and c) empirically-based simulations for the three models along with the comparisons with the model-based results.

4.1. Model-based Simulations with Set Trimming:

Table 4.1.1 summarizes the means and standard deviations of the item parameter estimates for each pool used in the simulations, as well as the mean and standard deviation of the estimated ability distribution. This table indicates that the items were significantly easier on average than the abilities of the candidates as estimated on the reference form (Form 38A).

Table 4.1.1: Item Parameter Estimates for 1-, 2-, & 3-PL Item Pools in Model-Based Simulations

			a-Parameters		*b*-Parameters		*c*-Parameters		Abilities–38A	
Model	#Items	#Sets	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1-PL	399	64	0.59	-	-0.83	0.95	0.00	-	0.06	0.75
2-PL	400	64	0.49	0.19	-1.07	1.23	0.00	-	0.00	1.02
3-PL	398	64	0.70	0.26	-0.33	1.17	0.28	0.08	-0.01	1.01

Figure 4.1.1 illustrates the disparity between ability distribution and distribution of difficulty parameters for the 1-PL model. This graph presents the items and abilities graphed on the same scale, with items on the left and abilities on the right. It can be seen that about 20 percent of the items had difficulties at or below –1.91 and that about half of the items had difficulties at or below –1.15; whereas, only about 6 to 7 percent of the estimated ability distribution was at or below –1.15.

Note also that for the 3-PL model, the difference between the mean *b*-parameter estimate and the mean of the ability distribution in Table 4.1.1 was smaller than for the other two IRT models. This is most likely because the difficulty estimates with the 3-PL model accounted for a non-zero lower asymptote. Note also that the mean *c*-parameter estimate with the 3-PL model was 0.28, which suggested that low ability candidates still had a reasonable probability of answering most items correctly. The 1-PL and 2-PL models did not account for guessing, and these models tended to estimate lower *b*-parameters to compensate.

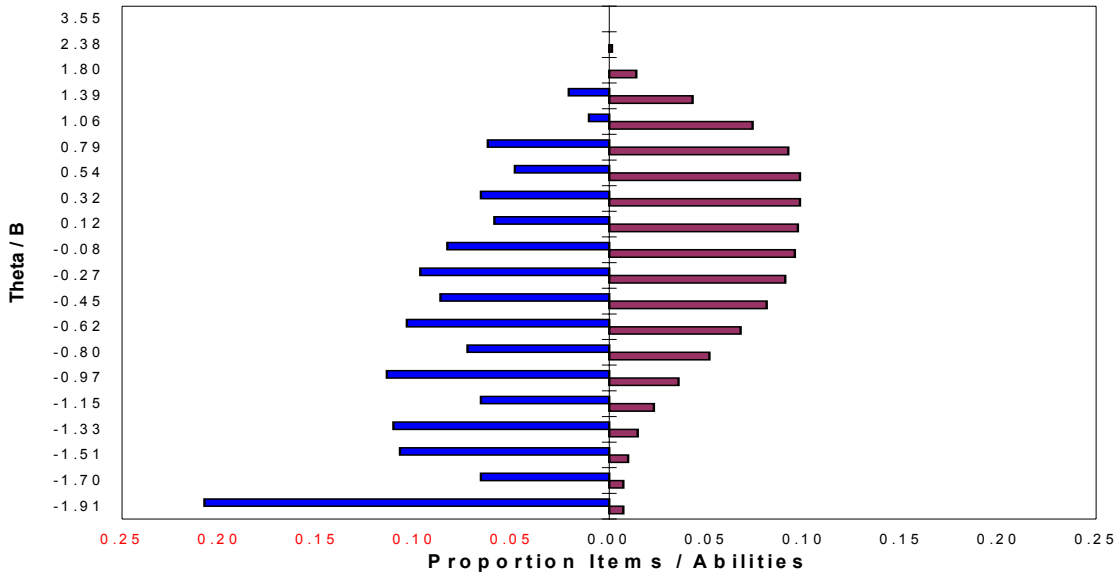

Figure 4.1.1: Estimated item difficulties and the estimated ability distribution for the 1-PL model.

The measurement precision of the simulated CAT examinations was summarized in terms of conditional standard errors of measurement (CSEMs), calculated at each generating number right-true score on the reference form (Form 38A), and in terms of overall simulation reliabilities. These results are presented in Table 4.1.2. and Figure 4.1.2. Although the true abilities differed across the 1-, 2-, and 3-PL models, because the generating true abilities across models corresponded to the same number right true score on the reference form, the CSEMs can be compared.

In addition, CSEMs based on the 1-, 2-, and 3-PL models were estimated at the same true number right true scores for the 55-item paper-and-pencil reference form itself, using the estimated item parameters for that form. These values are also presented in Figure 4.1.2. The overall simulation reliabilities were based on a weighted average of the CSEMs using the approach recommended by Green et al. (1984), and were calculated for the 32-item adaptive tests as well as for Form 38A based on each IRT model. These statistics are used routinely at ETS in evaluating adaptive test designs and item pool characteristics.

Table 4.1.2: CSEMs and Reliabilities for Model-Based Simulations with Set Trimming

	1-PL		2-PL		3-PL	
	CAT	P&P	CAT	P&P	CAT	P&P
Reliability	0.78	0.85	0.84	0.85	0.85	0.85

As shown in Table 4.1.2, the simulated reliabilities differed somewhat across models. Figure 4.1.2 indicates that the CSEMs for the 1-PL simulations were noticeably higher than those for the 2-PL and 3-PL simulations. This is mostly due to the fact that all items in the 1-PL simulations are assumed to be equally discriminating. In the 2-PL and 3-PL simulations, the more highly discriminating items are more likely to be chosen. Thus, the overall simulation reliabilities for the 2-PL and 3-PL based on 32-item adaptive tests were very similar to the estimated reliability for the 55-item paper-and-pencil test.

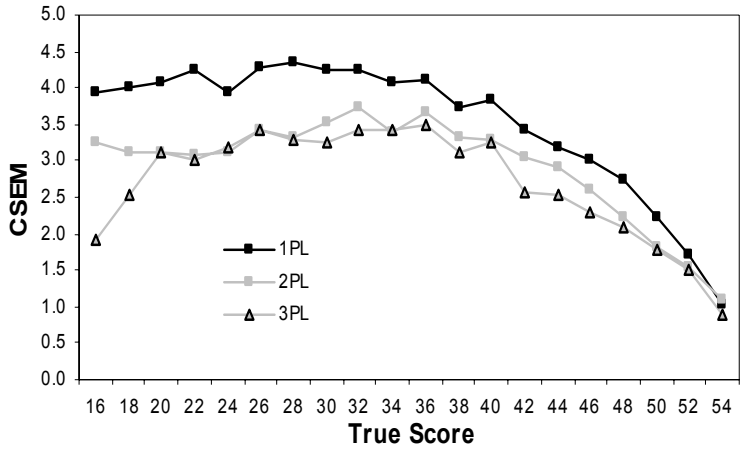

Figure 4.1.2: CSEMs for 1-, 2- and 3-PL CATs for model-based simulations.

Information about how well the content constraints were satisfied is provided in Table 4.1.3. This table lists the proportion of simulated cases where the number of items selected was either below the lower boundaries or above the upper boundaries shown in Table 3.2.1. It is clear that substantial content violations occurred for the cognitive categories in the simulations for all three models. This occurred because of the way that sets were selected for the simulations. Basically, the CAT algorithm first chooses a set on the basis of a single item within

the set that is considered the most desirable (this evaluation is based on a simultaneous consideration of content and statistical properties). However, once a set has been selected, the CAT algorithm is limited to selecting the remaining items in the set, which may not always include items in the cognitive category that is needed. This problem was exacerbated by the procedures followed to eliminate items from the sets that were extreme in difficulty compared to the other items.

Table 4.1.3: Content Constraint Violations for Model-Based Simulations

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 4.1.3 and 4.1.4 present cumulative frequencies of the stimulus exposure rates (upper graphs) and item exposure rates (lower graphs) based on the 1-, 2-, and 3-PL simulations. In these graphs, the desirable result is the one where the cumulative percentage of the pool used approaches 100 percent as quickly as possible. The patterns were very similar for the stimuli and items, which made sense given that most items within a set were used if the set was selected.

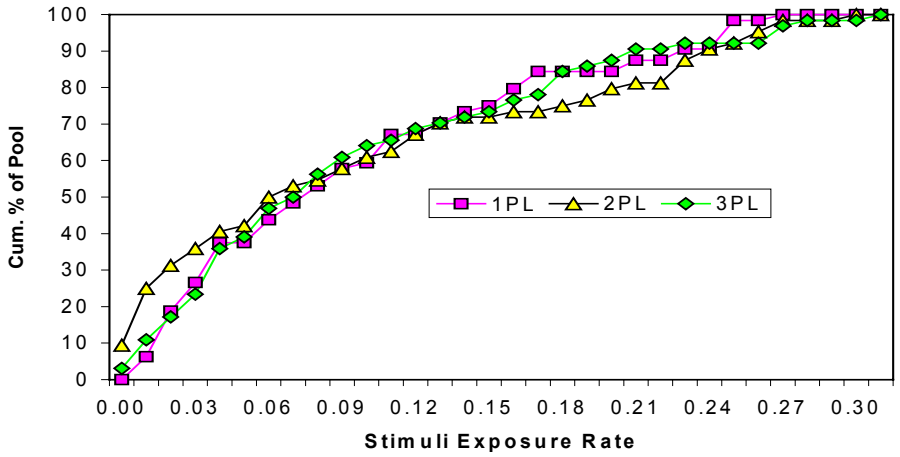

Figure 4.1.3: Cumulative exposure rates of stimuli for model-based Simulations.

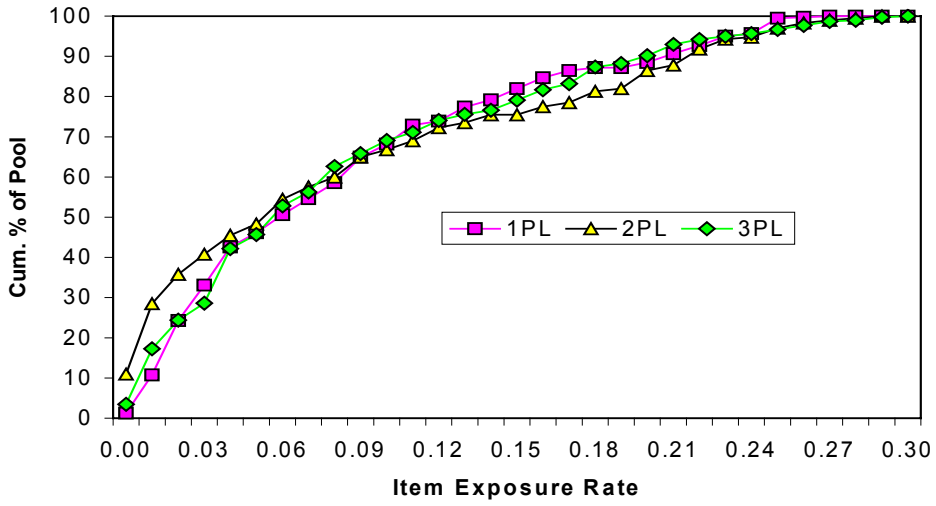

Figure 4.1.4: Cumulative exposure rates of set items for model-based simulations.

4.2. Model-Based Simulations without Set Trimming

In this section of the study, items that showed reasonably good model-data fit but were deleted from the previous simulation study because of disparate b-parameters within a set were recovered and simulations were repeated. For ease of comparison, Table 4.2.1 summarizes the means and standard deviations of the item parameter estimates for each pool used in the current

simulations, as well as in the previous simulations. The mean and standard deviation of the estimated ability distribution on the reference form (Form 38A) are provided also.

Table 4.2.1: Item Parameter Estimates for Item Pools for Simulations with and without Set Trimming

Pools	# Items	#Sets	*a*-Parameters		*b*-Parameters		*c*-Parameters		Abilities-38A	
			Mean	SD	Mean	SD	Mean	SD	Mean	SD
1-PL w/ trimming	399	64	0.59	-	-0.83	0.95	0.00	-	0.06	0.75
1-PL w/o trimming	434	64	0.59	-	-0.83	0.96	0.00	-		
2-PL w/ trimming	400	64	0.49	0.19	-1.07	1.23	0.00	-	0.00	1.02
2-PL w/o trimming	438	64	0.48	0.18	-1.04	1.26	0.00	-		
3-PL w/ trimming	398	64	0.7	0.26	-0.33	1.17	0.28	0.08	-0.01	1.01
3-PL w/o trimming	438	64	0.7	0.26	-0.31	1.19	0.28	0.08		

Table 4.2.1 indicates that there was very little difference between item pools with set trimming and those without set trimming with regards to item parameter means and standard deviations. Items in all pools were significantly easier on average than the abilities of the candidates as estimated on the reference form.

Measurement precision, in terms of overall simulation reliabilities, of the CAT examinations simulated from the item pools with and without set trimming are summarized and presented in Table 4.2.2 and Figures 4.2.1 to 4.2.3. Measurement precision results from previous simulations are also displayed for comparison. Similar to what was observed in the previous study, the CSEMs for the 1-PL simulations were higher than those for the 2-PL and 3-PL simulations. The overall reliabilities of 2-PL and 3-PL CAT simulated exams were close to the P&P reference test, but the reliability of 1-PL CAT exam was relatively low compared to both its 2-PL and 3-PL counterparts and the P&P reference form.

Table 4.2.2: Reliabilites for Model-Based Simulations (with/without set trimming)

	1-PL			2-PL			3-PL		
	CAT w/trim	CAT w/o trim	P&P	CAT w/trim	CAT w/o trim	P&P	CAT w/trim	CAT w/o trim	P&P
Reliability	0.78	0.78	0.85	0.84	0.83	0.85	0.85	0.84	0.85

Furthermore, the plots show that except for the 2-PL situation where CSEMs for no trimming were relatively higher than those for trimming at low levels of true scores, the discrepancies of CSEMs between trimming and non-trimming situations were miniscule. Although the purpose of set trimming was to enhance measurement precision, the gain was not obvious in our simulation studies except for a slight increase in reliability for 2-PL and 3-PL situations. This occurred because the number of deleted items in previous set trimming was not large per se, hence its influence was trivial. Another reason is that although these trimmed items were recovered, the probabilities of their getting selected under the condition that the sets they are in were selected would be low due to their extreme difficulties, as would be determined by the CAT algorithm. There might be situations when these items were considered the most desirable when the decision of set selection was about to be made, in which case a disparate difficulty within set would not prevent them from getting selected, but that probability is small too. If few of the items previously trimmed were administered, trimming or not trimming would not make much difference with respect to CSEMs and overall reliability.

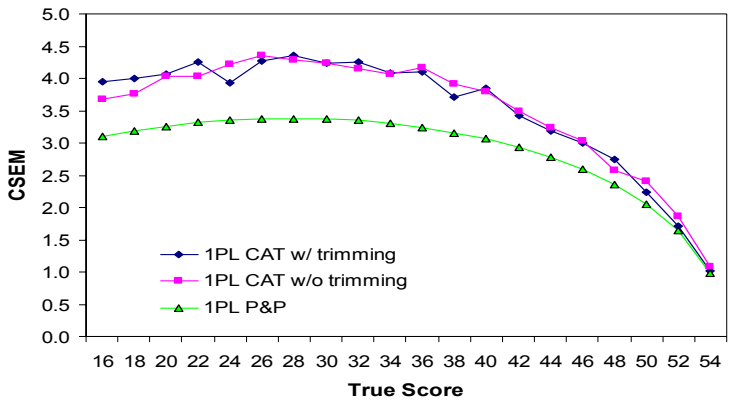

Figure 4.2.1: CSEMs for 1-PL P&P and CATs for model-based simulations (with and without set trimming).

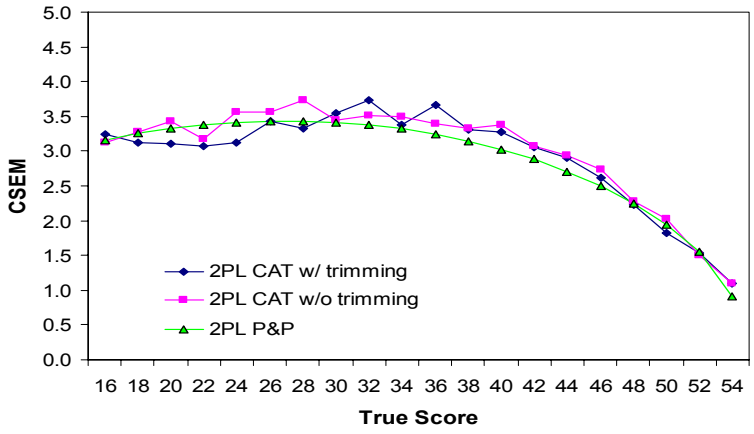

Figure 4.2.2: CSEMs for 2-PL P&P and CATs for model-based simulations (with and without set trimming).

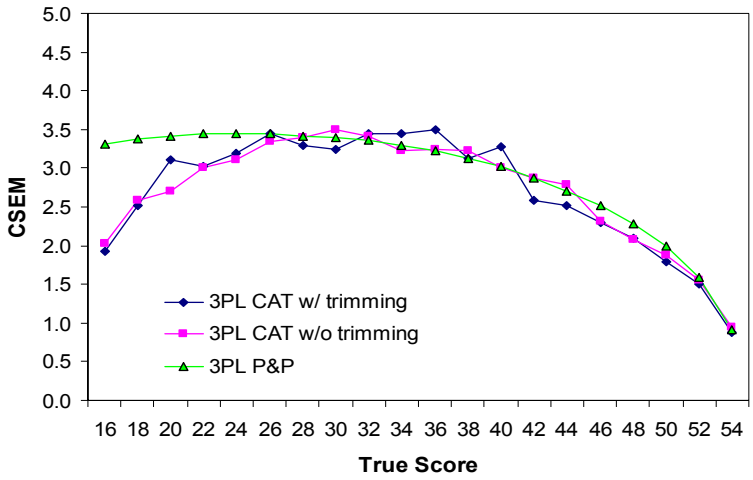

Figure 4.2.3: CSEMs for 3-PL P&P and CATs for model-based simulations (with and without set trimming).

The absolute bias in ability estimation for the trimmed and non-trimmed cases across the three models is presented in Figures 4.2.4. to 4.2.6. The figures show that the bias for 1-PL was largest across most of the ability levels. The difference between trimming and not trimming of the sets was apparent most in the 3-PL model where bias was greatly reduced by trimming the sets.

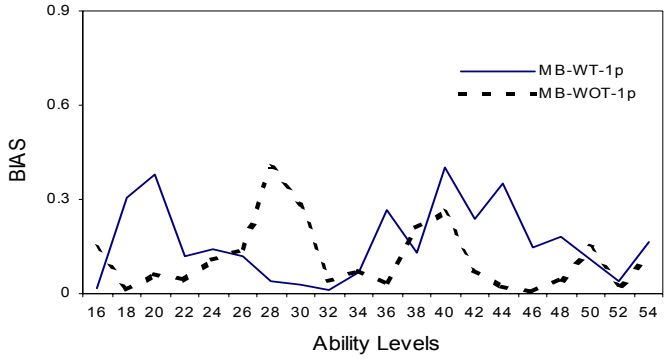

Figure 4.2.4: Bias for 1-PL CATs for model-based simulations (with/without set trimming).

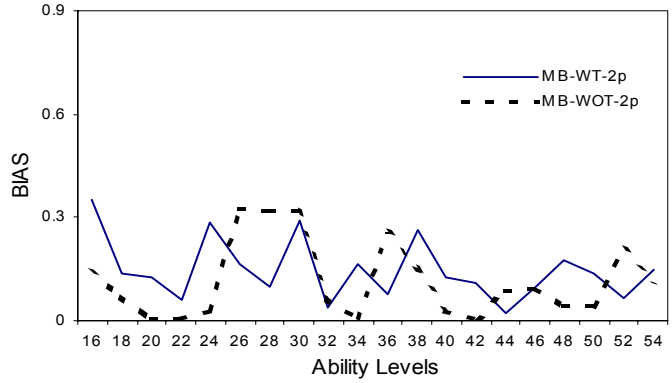

Figure 4.2.5: Bias for 2-PL CATs for model-based simulations (with/without set trimming).

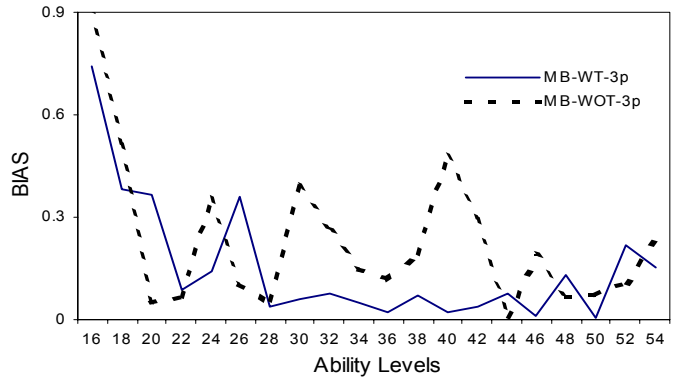

Figure 4.2.6: Bias for 3-PL CATs for model-based simulations (with/without set trimming).

Information about how well the content constraints were satisfied is provided in table 4.2.3. This table lists the proportion of simulated cases where the numbers of items selected

were either below the lower boundaries or above the upper boundaries. Minimum and maximum numbers of items in each content category selected were also presented. Content constraint violations from initial simulation with set trimming are also provided for comparison purpose.

As expected, not trimming increased the proportion of cases for which the desired number of items across the cognitive categories was selected, and improvement was most noticeable in the "Natural Science" content category. Nevertheless, Table 4.2.3 indicates that there were still considerable content violations for some item cognitive constraints across all three models.

Table 4.2.3: Content Constraint Violations for Model-Based Simulations (with and without trimming

Content Constraint	Targeted #Items			%Violations*			Min.Adm.*			Max.Adm.*		
	Low	High	Wt.	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 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: NatSci	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: SocSci	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: Six	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: Five	4	4	10	0 0	0 0	0 0	4 4	4 4	4 4	4 4	4 4	4 4
I: Comp	8	12	90	.08 .08	.04 .05	.05 .08	5 6	6 5	4 5	16 16	17 14	14 16
I: Eval	4	8	90	.03 .02	.02 .02	.02 .03	3 3	3 3	3 3	10 10	11 9	11 10
I: Appl	7	10	90	.07 .12	.07 .08	.07 .09	4 5	3 5	5 4	14 13	13 12	12 12
I: Incorp	6	9	90	.06 .08	.06 .12	.14 .12	4 3	4 4	3 4	11 12	10 12	10 12
I: Human	10	12	10	0 0	0 0	0 0	10 10	10 10	10 10	12 12	12 12	12 12
I: NatSci	10	11	10	0 .11	0 .34	0 .38	10 10	10 10	10 10	11 12	11 12	11 12
I: SocSci	10	12	10	0 0	0 0	0 0	10 10	10 10	10 10	12 12	12 12	12 12

* Numbers in the first line of cells represent non-trimming scenario. Numbers in the second line of cells represent set trimming scenario.

Figure 4.2.7 presents cumulative frequencies of the stimulus exposure rates and Figure 4.2.8 presents cumulative frequencies of the item exposure rates for the 1-PL, 2-PL, and 3-PL simulations. The patterns of exposure rate frequencies were very similar across the three IRT models.

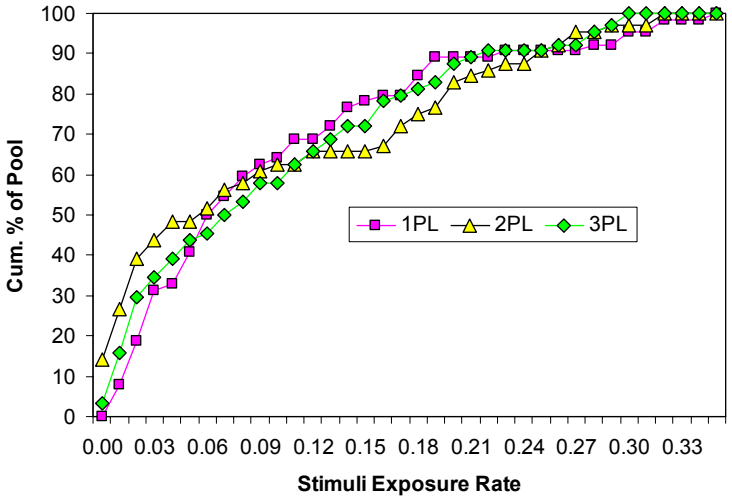

Figure 4.2.7: Cumulative exposure rates of stimuli for model-based Simulations (without trimming).

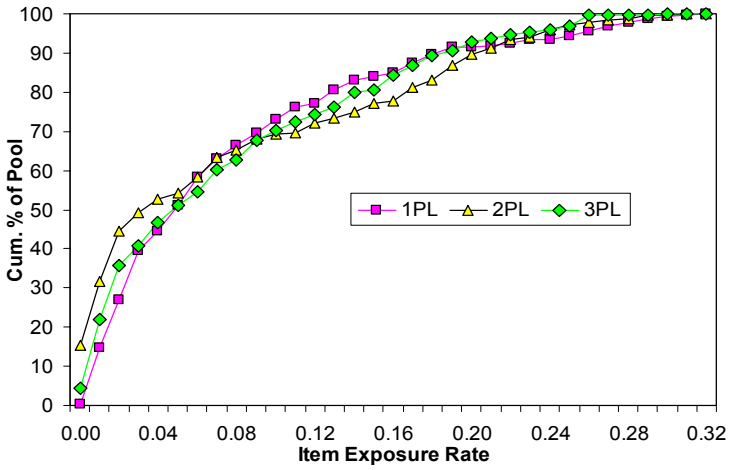

Figure 4.2.8: Cumulative exposure rates of set items for model-based simulations (without trimming).

To examine the influence of trimming and no trimming on exposure rates, plots of pair-wise comparisons were displayed. Figures 4.2.9 to 4.2.11 compare exposure rates of VR stimuli between trimming and non-trimming situations for 1-, 2- and 3-PL respectively. Figure 4.2.12 to

4.2.14 compare exposure rates of VR items between trimming and non-trimming situations for 1-, 2- and 3-PL respectively. These figures show that the difference between trimming and non-trimming was negligible with respect to exposure rates. Although trimming seemed to outperform non-trimming scenario in stimuli exposure rates for 1-PL, generally they exhibited very similar results.

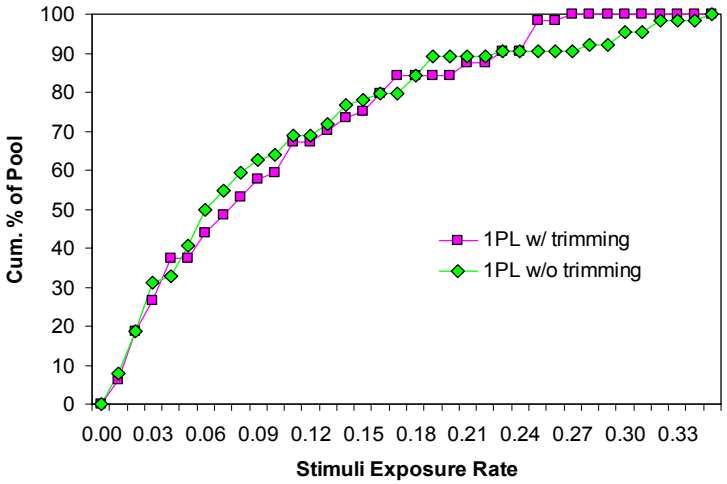

Figure 4.2.9: Cumulative exposure rates of stimuli for 1-PL for model-based simulations (with and without trimming).

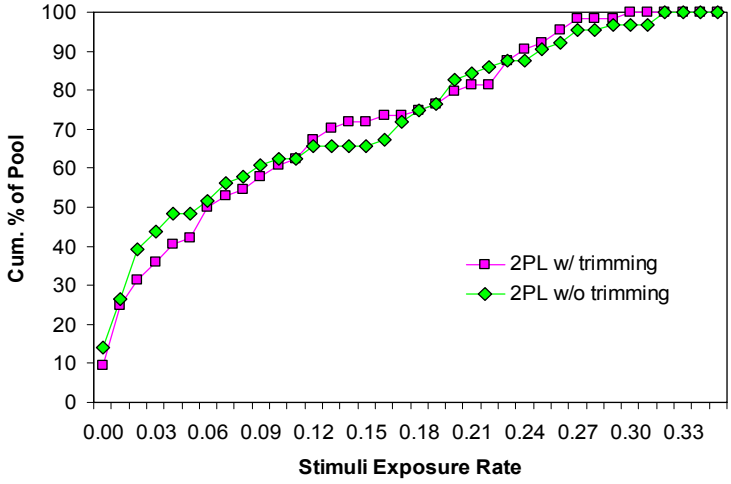

Figure 4.2.10: Cumulative exposure rates of stimuli for 2-PL for model-based simulations (with and without trimming).

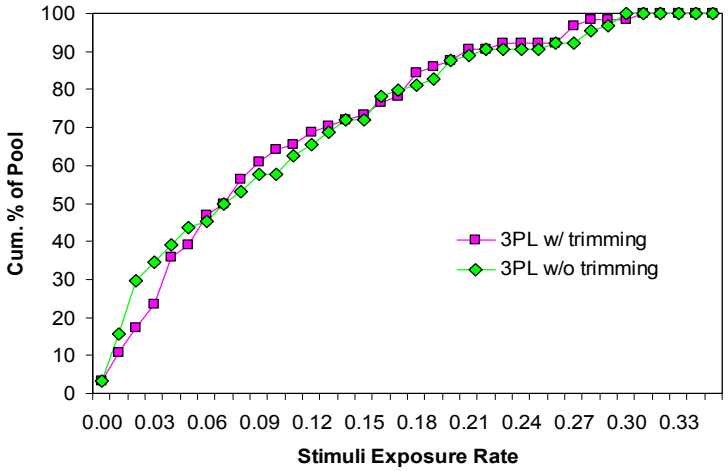

Figure 4.2.11: Cumulative exposure rates of stimuli for 3-PL for model-based simulations (with and without trimming).

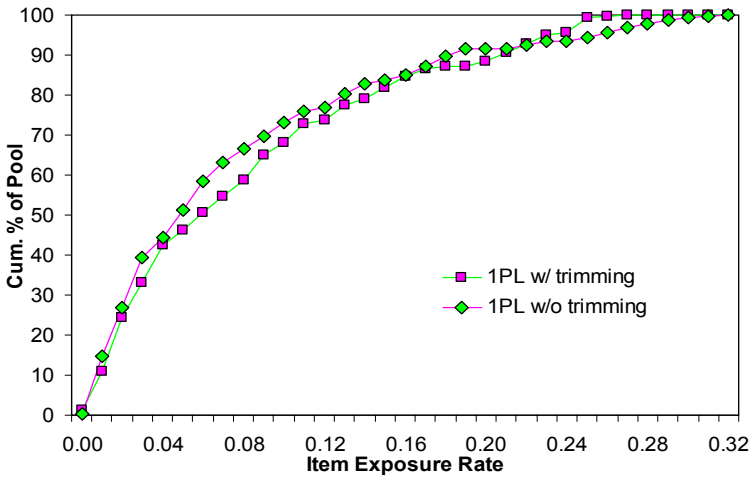

Figure 4.2.12: Cumulative exposure rates of set items for 1-PL for model-based simulations (with and without trimming).

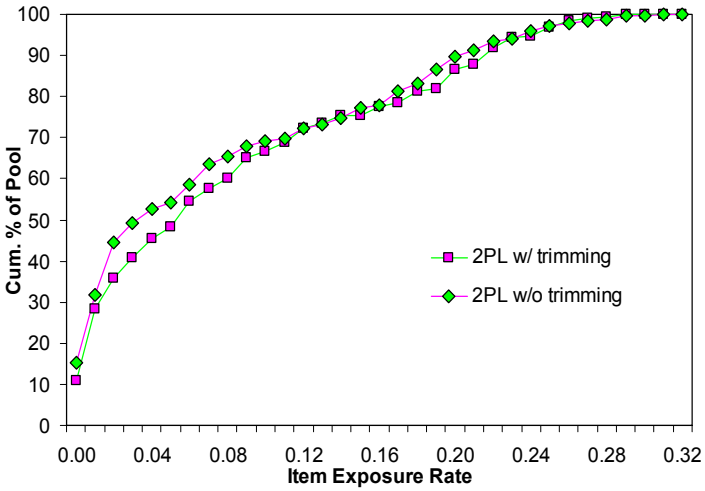

Figure 4.2.13: Cumulative exposure rates of set items for 2-PL for model-based simulations (with and without trimming).

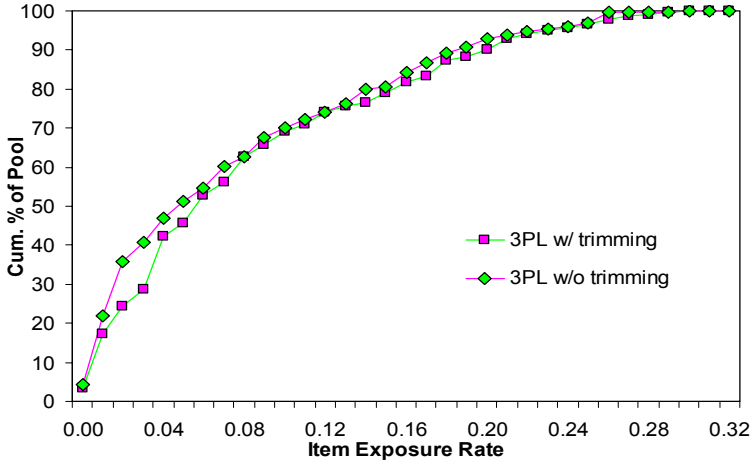

Figure 4.2.14: Cumulative exposure rates of set items for 3-PL for model-based simulations (with and without trimming).

In summary, simulations without set trimming gave quite similar results to simulations with trimming. Trimming seemed to give better results than non-trimming with respect to measurement precision and exposure rates, but only to a very small degree. On the other hand, however, non-trimming resulted in fewer violations of content constraints, which would be attractive from the content point of view.

4.3. Empirically-Based Simulations without Set Trimming

In previous studies, it was found that simulations without set trimming gave quite similar results as simulations with trimming, except that non-trimming performed slightly better than trimming in that it resulted in fewer violations of content constraints. Therefore, the simulations carried out in this part of the study only used the item pool without set trimming. Results from this study and the previous part of the study using model-based data generation without set trimming are compared and summarized below. For simplification, titles for some of the graphs and tables use 'MB' as referring to model-based data generation without trimming, and 'EB' as referring to empirically based data generation without trimming.

Measurement precision, in terms of CSEMs and overall simulation reliabilities, is summarized and presented in Table 4.3.1 and Figures 4.3.1 to 4.3.3. Similar to what was observed in model-based simulations, the CSEMs for the 1-PL empirically-based simulations were higher than those for the 2-PL and 3-PL simulations. The overall reliabilities for 2-PL and 3-PL CAT simulated exams, for both empirically based and model based situations, were very close to the P&P reference test, but the reliabilities of 1-PL CAT exams were relatively low compared to their 2-PL and 3-PL counterparts and the P&P reference form.

Table 4.3.1: Reliabilities for Model-Based and Empirically-Based Simulations (without set trimming)

	1-PL			2-PL			3-PL		
	EB	MB	P&P	EB	MB	P&P	EB	MB	P&P
Reliability	0.77	0.78	0.85	0.83	0.83	0.85	0.84	0.84	0.85

The results also showed that for 1-PL model, in comparison with model-based simulations, reliability in empirically based simulation were slightly lower, and CSEMs were considerably higher at the lower end of the ability scale. This made sense, as model data misfit is not unusual in applications of the 1-PL model. As shown in Figure 2.1, a low ability examinee was over-predicted by the model, and a high ability examinee, under-predicted. The results suggested that in the simulation process low ability examinees often received items with difficulties above their ability levels. Since data were generated from model-based

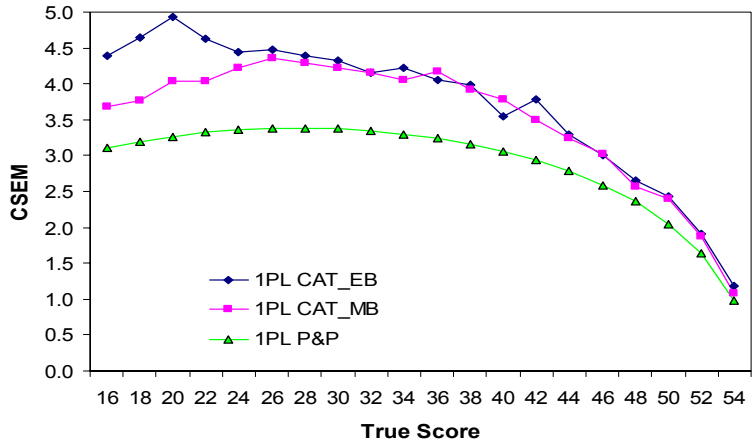

Figure 4.3.1: CSEMs for 1-PL P&P and CATs for model-based and empirically-based simulations (without set trimming).

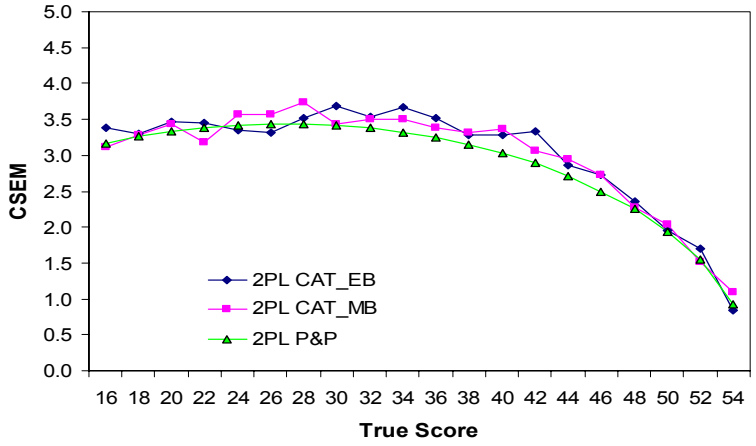

Figure 4.3.2: CSEMs for 2-PL P&P and CATs for model-based and empirically-based simulations (without set trimming).

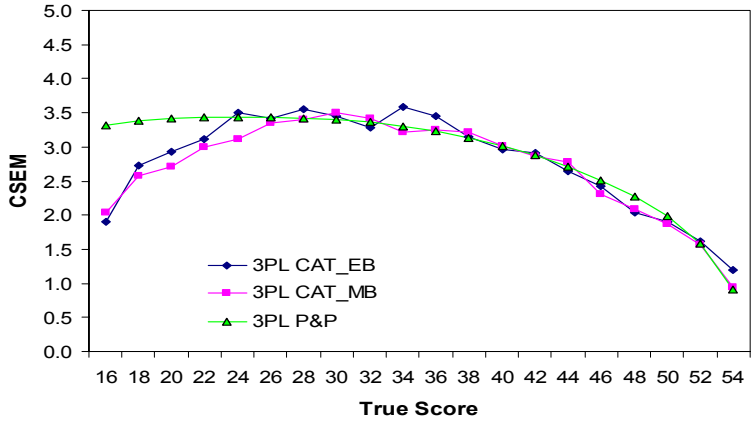

Figure 4.3.3: CSEMs for 3-PL P&P and CATs for model-based and empirically-based simulations (without set trimming).

probability that ignored the model-data misfit at low ability levels, CSEMs from model-based simulation at lower ability levels were smaller. Since 2-PL and 3-PL models provided fairly good model-data fit for most of the MCAT items, the discrepancy between the model-based and empirically based probabilities was small throughout the ability scale for those models. As a result, the two simulations give very similar results.

The absolute bias in ability estimation for the trimmed and non-trimmed cases across the three models is presented in Figures 4.3.4 to 4.3.6. The figures show that the bias for 1-PL was largely underestimated when the model-based data generation was used to simulate examinee responses. The significantly large bias at the lower end of the ability scale was caused by the interaction between the limitation of the maximum likelihood estimation algorithm and model-data misfit during calibrations.

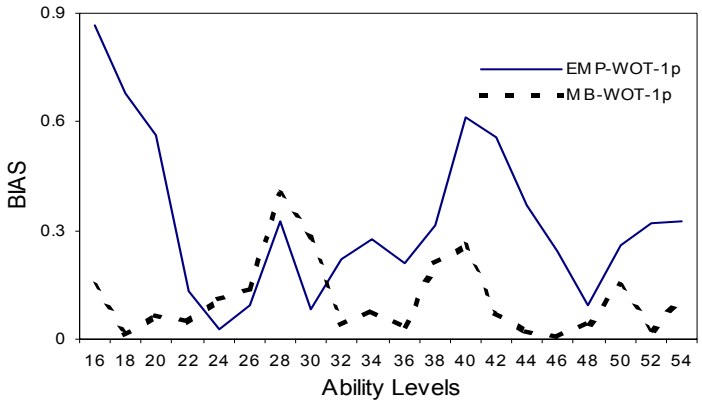

Figure 4.3.4: Bias for 1-PL CATs for model-based and empirically-based simulations.

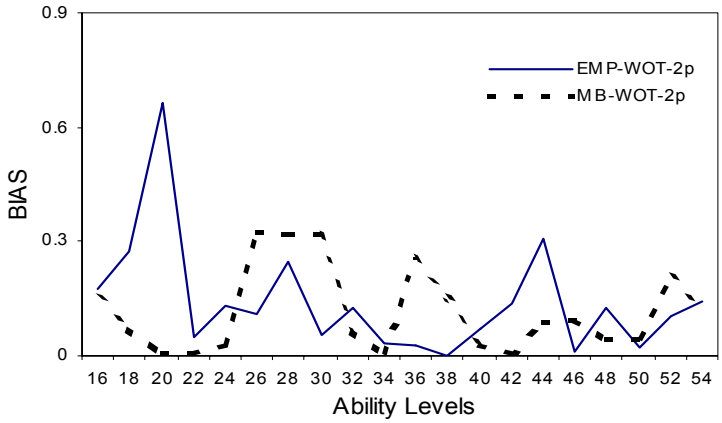

Figure 4.3.5: Bias for 2-PL CATs for model-based and empirically-based simulations.

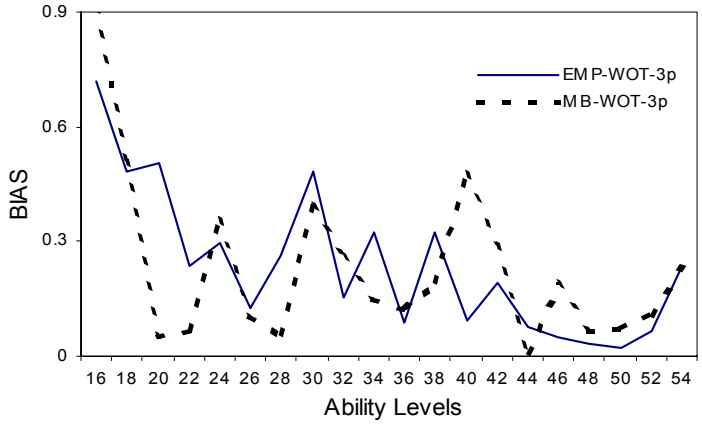

Figure 4.3.6: Bias for 3-PL CATs for model-based and empirically-based simulations.

Information about how well the content constraints were satisfied in empirically based simulations as well as model-based simulations is provided in Table 4.3.2. This table lists the proportion of simulated cases where the numbers of items selected were either below the lower boundaries or above the upper boundaries. Minimum and maximum numbers of items in each content category selected were also presented.

For empirically based and model-based CAT simulations, the degree of content constraint violations was very much alike. This makes sense because changing the probability for data generation should not directly affect the satisfaction of content specifications in item selection.

Table 4.3.2: Content Constraint Violations for Model-Based and Empirically-Based Simulations (without set trimming)

Content Constraint	Targeted #Items			%Violations*			Min.Adm.*			Max.Adm.*		
	Low	High	Wt	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 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: NatSci	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: SocSci	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: Six	2	2	10	0 0	0 0	0 0	2 2	2 2	2 2	2 2	2 2	2 2
S: Five	4	4	10	0 0	0 0	0 0	4 4	4 4	4 4	4 4	4 4	4 4
I: Comp	8	12	90	.08 .08	.04 .04	.05 .05	5 5	6 6	5 4	16 16	17 17	14 14
I: Eval	4	8	90	.03 .03	.02 .02	.02 .02	2 3	3 3	3 3	10 10	11 11	11 11
I: Appl	7	10	90	.07 .07	.08 .07	.06 .07	5 4	3 3	5 5	13 14	13 13	13 12
I: Incorp	6	9	90	.07 .06	.06 .06	.15 .14	4 4	4 4	4 3	11 11	10 10	10 10
I: Human	10	12	10	0 0	0 0	0 0	10 10	10 10	10 10	12 12	12 12	12 12
I: NatSci	10	11	10	0 0	0 0	0 0	10 10	10 10	10 10	11 11	11 11	11 11
I: SocSci	10	12	10	0 0	0 0	0 0	10 10	10 10	10 10	12 12	12 12	12 12

* Numbers in the first line of cells are results from the empirically based CAT simulation study. Numbers in the second line of cells are results from the model based CAT simulation study.

Figure 4.3.7 presents cumulative frequencies of the stimulus exposure rates and Figure 4.3.8 presents cumulative frequencies of the item exposure rates based on empirically-based and model based simulations. Again, as with previous simulation studies, the patterns of exposure rate frequencies were very similar across 1-, 2- and 3-PL models in empirically based simulations. The plots also showed that there was almost no difference between empirically-based and model-based simulations with respect to item exposure rates. For stimuli exposure rates, there was some difference, but very small. This is again, not unexpected, as changing the probability for data generation should not affect exposure rate control in CAT simulations.

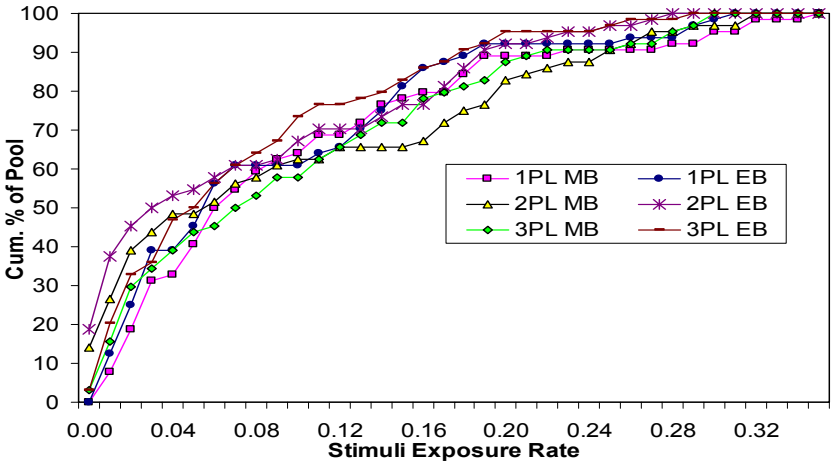

Figure 4.3.7. Cumulative exposure rates of stimuli for model-based and empirically-based simulations (without trimming).

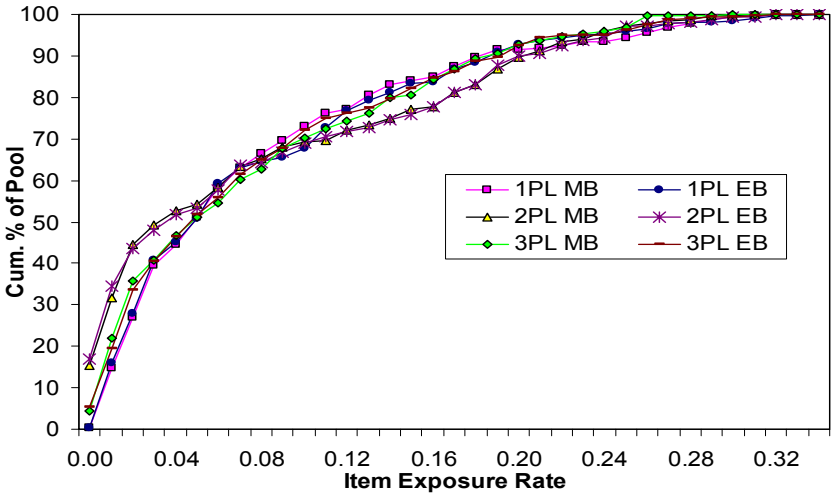

Figure 4.3.8. Cumulative exposure rates of set items for model-based and empirically-based simulations (without trimming).

In conclusion, empirically-based simulations gave similar results to model-based simulations when the model applied fitted the data well. When there was considerable amount of model data misfit, the empirically-based simulations gave more realistic results than model based simulations.

5. Summary

The CAT simulations that were carried out suggested that measurement precision equivalent to the current paper-and-pencil MCAT Verbal Reasoning test could be achieved with a 32-item adaptive test based on the 2-PL or 3-PL models. However, simulations of 32-item adaptive tests based on the 1-PL model were far less reliable. These findings were expected, given that the 2-PL and 3-PL models make use of the differential discriminating power of items in constructing and scoring adaptive tests. Although the 2-PL and 3-PL simulations made slightly less uniform use of the item pools, the differences between these models and the 1-PL model were surprisingly small.

In terms of content coverage, the set-based nature of the Verbal Reasoning test and the irregular representation of items with different cognitive categories across different passages made it all but impossible to consistently satisfy targeted test specifications. This issue will have to be addressed if the implementation of an MCAT using CAT or adaptive testlets is to be seriously contemplated.

The results showed that when there was considerable amount of model data misfit, the model-based simulations gave smaller CSEMs at certain ability levels, which are misleading. The empirically-based simulations provided a more reliable way of evaluating a CAT design before it's implementation.

An administration of a 2-PL CAT with reliability comparable to P&P reliability using almost half the P&P test length is a very positive finding of the study. The use of a 2-PL instead of a 3-PL model is recommended because of the simplicity of the 2-PL model.

References

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 41, 443–459.

Davey, T. (2002). Personal Communication. Educational Testing Service, Princeton, NJ.

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

Muraki, E., & Bock, R. D. (1991). PARSCALE: Parameter scaling of rating data [Computer software]. Chicago, IL: Scientific Software.

Stocking, M. L., & Lewis, C. (2000). Methods of controlling the exposure of items in CAT. In W. J. van der Linden & C. A. W. Glas (Eds.), Computer adaptive testing: Theory and practice (pp. 163–182). Norwell, MA: Kluwer Academic Publishers.

Stocking, M. L., & Swanson, L. (1993). A model and heuristic for solving very large item selection problems. Applied Psychological Measurement, 17, 151-66.