# The Effects of Model Misfit in Computerized Classification Test

Hong Jiao
Florida State University
hjiao@usa.net

Allen C. Lau
Harcourt Educational Measurement
allen_lau@harcourt.com

Paper presented at the annual meeting of the National Council of Educational Measurement, Chicago, IL, April 2003.

Abstract

This study investigated the effects of model misfit on classification decisions made from computerized classification test using sequential probability ratio test. One of the three unidimensional binary IRT model (1PL, 2PL, and 3PL) was assumed the true model and the other models were treated as the misfit models. The results indicated that when the 1PL or the 2PL IRT model was the true model, using either of the other two IRT models did not affect the classification decisions. However, when the 3PL IRT model was the true model, using the 1PL IRT model increased the false positive classification error rates to the level much larger than its nominal level, while using the 2PL IRT model in CCT increased the false negative classification error rates to the level above the nominal level. Thus, it was concluded that IRT model selection and evaluation is highly recommended in computerized classification test using sequential probability ratio test.

# The Effects of Model Misfit in Computerized Classification Test

## Introduction

Item Response Theory (IRT) plays an important role in Computerized Classification Test (CCT) using Sequential Probability Ratio Test (SPRT). It facilitates item pool calibration, item selection, classification decision-making, and test termination. The implementation of an IRT based CCT using SPRT relies on the invariance nature of the item parameter estimation using an IRT model. The invariance of the item parameter estimates holds in proportion to the extent that the assumptions of the IRT model are met and the model fits the data (Hambleton & Swaminathan, 1985; Embretson & Reise, 2000).

This study was to investigate extensively the effect of model misfit on the mastery decisions made from the IRT-based CCT using SPRT under various testing conditions. A series of CCTs were simulated to make dichotomous classification decisions based on dichotomous item response data. The three unidimensional IRT models, the one-parameter logistic (1PL), the two-parameter logistic (2PL), and the three-parameter logistic (3PL) IRT models were specified as the true model respectively and the true model were compared with the misfit models in terms of classification accuracy, error rates, and the average test length under different study conditions.

## Background

A mastery test is used to classify the examinees into one of two categories: mastery (pass) or non-mastery (fail). Certification or licensure testing is a good example of this. An

IRT-based CCT using SPRT is one of the methods used to serve this purpose. Such a test starts with administering an item with the largest amount of item information at the cut-score point. Then, the likelihood ratio is computed and compared with the lower and upper boundaries determined by the nominal type I and type II error rates. If it is equal to or larger than the upper bound, the examinee is classified as a master. If it is equal to or smaller than the lower bound, the examinee is classified as a non-master. If it is between the two boundaries, test continues by administering the next item with the most item information at the cut-score point.

Several factors can influence decision-making in a CCT using SPRT (Lau, 1996; Spray et al., 1997; Lau & Wang, 1998, 1999, 2000; Kalohn & Spray, 1999; Lin & Spray, 2000; & Parshall et al., 2002). These factors include the location of the cut score, the width of the indifference region, test length, item pool characteristics, content balancing, and item-exposure control rate. The cut score and the width of the indifference region affect the likelihood ratio. Test length is a function of the width of indifference region (Lau, 1996). Item pool characteristics affect the test length. If an item pool consists of more items with the maximum information near the cut score point, fewer items will be used. To balance complex content of a test, more items are needed. A randomization scheme (Lin & Spray, 2002) can be used to control item exposure rate. A stratum depth of 1 means no exposure control. Less item exposure control, fewer items needed.

In testing, some practitioners recommended using only one IRT model as pointed out by Kalohn & Spray (1999). Some people (Embretson & Reise, 2000) thought that the model choice in applying IRT models is obvious. By contrast, some other people thought that the model selection and evaluation is a central issue in the IRT applications. Owing to the significance of the IRT model in CCT using SPRT and the controversy over the model selection and evaluation in the IRT application, it is necessary to find empirical evidence related to the effect of model misfit on classification decisions in CCT using SPRT.

Lau (1996) and Spray et al. (1997) examined the robustness of IRT-based CCT using SPRT when unidimensionality assumption was violated. Their studies indicated that CCT using SPRT was robust to the violation of unidimensionality assumption. Though Kalohn & Spray (1999) studied the effect of model misspecification on classification decisions based on CCT, they only adopted 3-PL IRT model as the true model and compared the decision differences between the true model and 1-PL IRT model under the variation of test length and item exposure rate control. This issue awaits further exploration.

## The Purpose of the Study

The purpose of this study was to explore more thoroughly the effect of model misfit on classification decisions made from an IRT based CCT using SPRT. The 1-PL, the 2-PL, and the 3-PL IRT models were specified as the true model respectively and the true model were compared with the misfit models in terms of error rates, classification accuracy, and the average test length under different simulated study conditions. More specifically, this study intended to answer the following research questions.

First, when the 1-PL IRT model was the true model, if the 3-PL or the 2-PL IRT model was used in CCT, what were the differences in terms of classification accuracy, error rates, and the average test length?

Second, when the 2-PL IRT model was the true model, what were the changes in light of classification accuracy, error rates, and the average test length in the 1PL and the 3PL-based CCT?

Third, when the 3-PL IRT model was the true model, would the use of the 1PL and the 2PL IRT model seriously affect the CCT results?

## Methods

### Procedure

This study made use of the Monte Carlo simulation technique. It consisted of three sub-studies. Study 1, 2, and 3 assumed the 1-PL, the 2-PL, and the 3-PL IRT model as the true model respectively. In each sub-study, the true model was used to generate dichotomous item responses, given the true item parameters and the true ability parameters. Then, based on the generated item responses, the item pool was calibrated using the misfit models. After the CCT using SPRT procedure was simulated, the mastery decisions based on the two misfit IRT models were compared with those based on the true model. For instance, in Study 1, the 1-PL IRT model was the true model. Item responses were generated based on the known item parameters and the true ability parameters using the 1-PL IRT model. Based on the generated item responses, the item pool was calibrated using the 2-PL and the 3-PL IRT models respectively. Then, the CCT using SPRT was simulated using the item parameters from the 1-PL (true model), the 2-PL (misfit model), and the 3-PL (misfit model) correspondingly. The mastery decisions based on the 1-PL IRT model, those based on the 2-PL, and the 3-PL IRT models were then compared. All CCTs were simulated using S-PLUS. Table 1 illustrates the research design.

[Insert Table 1 about here]

In all simulated CCTs, items were selected according to the Fisher item information at the cut-score point. The true cut score was 0 on the latent ability scale. The width of the indifference region for the SPRT was fixed as 0.5. The nominal type I and type II error rates were both set at the 0.05 level. No content-balancing was imposed.

In each sub-study, simulation was implemented under four study conditions that were the combination of two manipulated factors. The two factors were the constraints of the test length and the stratum depth indicating the item exposure rate control in the randomization scheme. For one scenario, the minimum and the maximum test length were set as 60 and 120 items respectively to represent a realistic testing situation. Accordingly, the item pool size was calculated to be 300 to meet the requirement of 20% maximum item exposure control rate based on Parshall et al. (2002, P155). The levels of the two manipulated factors were specified as follows:

1. Test length:
   - i). minimum=60; maximum=120
   - ii). minimum=1; maximum=300
2. Stratum depth (item exposure control):
   - i). 1
   - ii). 5

The four study conditions resulted, i.e., test conditions with no constraints (F1), with only test length constraint (S1), with only item exposure constraint (F5), and with both constraints (S5).

<u>Data Simulation</u>

In each sub-study, the true ability and the true item parameters were preset to generate item responses. The true ability parameters were generated from a normal distribution with mean of 0 and standard deviation of 1 for 10,000 simulated examinees. They remained the same for all three sub-studies. The true item difficulty parameters were the same for all three sub-studies and generated from a normal distribution with mean of 0 and standard deviation of 1 within a range from –1.5 to +1.5. The discriminating parameters for Study 2 and 3 were the same and generated from a lognormal distribution within the range from 0.4 to 1.6. Pseudo-guessing parameters in Study 3 were fixed at 0.2.

The examinees' dichotomous responses were simulated as follows. First, the probability of a person correctly answering a selected item was obtained by incorporating the true item and the true ability parameters into the true IRT model. This probability was then compared with a random number generated from a uniform distribution from 0 to 1. If the probability was larger than or equal to the random number, a correct response of 1 was obtained, otherwise, an incorrect response of 0 resulted.

The item parameters for the misfit model based CCTs were obtained by calibrating the true item pool using the misfit model and the item response generated based on the true model and the true item parameters. A sample of 2,000 examinees was randomly selected from the 10,000 examinees. The item responses of the 2000 examinees to the 300 items in each true item pool were generated using S-PLUS. The items were calibrated using BILOG-MG.

Evaluation Criteria

To examine the model misfit effect in an IRT-based CCT using SPRT, the mastery decisions from the true and the misfit IRT models were compared in terms of four evaluation criteria. They were correct classification rates, Type I error rates, Type II error rates, and average test length used for classification decisions. The classification accuracy is the rate of consistent classification of examinees based on the simulated CCTs and the examinees' true class category. The Type I error rates refers to the false positive classification error rates while the Type II error rates means the false negative classification error rates. The average test length is the average number of items used for classification decisions for each examinee. It is an indicator of cost-effectiveness.

# Results

The main purpose of this research was to study the effect of model misfit on the IRT based CCT using SPRT. As the misfit model based CCT used item parameters calibrated using the misfit model and the item response generated from the true model and the true item parameters, the effect of item parameter estimation error should be separated from the effect of model misfit in CCT using SPRT. Based on Spray & Reckase (1987) and Kalohn & Spray (1999), the measurement error in the item parameter estimation did not affect the CCT classification results. Thus, CCT comparisons were only conducted between the true models and the misfit models. This section presents the results from the comparisons between the true model and the misfit model based CCTs. Table 2 summarizes the descriptive statistics of the generated true item parameters. Tables 3 to Table 8 summarize the descriptive statistics for the calibrated item parameters in each sub-study. Table 9 shows the results for Study 1. Study 2 results are presented in Table 10. Table 4 displays the results for Study 11.

[Insert Table 2 about here]

[Insert Table 3 about here]

[Insert Table 4 about here]

[Insert Table 5 about here]

[Insert Table 6 about here]

[Insert Table 7 about here]

[Insert Table 8 about here]

## The Effect of Misfit Models

In Study 1, the 2PL-based CCT using the calibrated item parameters had almost the same correct classification rates as those based on the true model. The differences were smaller than

1%. The false positive and the false negative error rates changed slightly. None of the error rates was larger than their nominal value of .05. The average test length used for classification changed little.

When the 3PL was used instead of the true model-1PL to make classification decisions, the percentage of correct classification decreased slightly. The biggest decrease was 1%. The false positive error rates increased evidently with the smallest amount of 1.7%. On the contrary, the false negative error rates decreased with the largest decrease of 1.4%. However, none of the error rates was larger than .05. On average, fewer items were needed when the 3PL IRT model was used in classification decisions when the 1PL IRT model was the true model.

[Insert Table 9 about here]

In Study 2, the 2PL IRT model was the true model. The correct classification rates were about 1% lower in the 1PL-based CCTs than the true model-2PL based CCTs. Compared to the true model based CCTs, the false positive error rates were higher in the 1PL-based CCTs with the largest discrepancy of 1.8% while the false negative error rates were lower in the 1PL-based CCTs. All the error rates were smaller than 5%. Generally, longer average test length was needed for the 1PL-based CCTs.

In Study 2, when the 3PL IRT model was used instead of the true model, 2PL, the correct classification rates decreased in most test conditions. The growth in the false positive error rates was noticeable. The false negative error rates showed the opposite trend. Neither of these two Types of error rates was larger than the nominal level. The average test length was almost the same in most cases.

[Insert Table 10 about here]

In Study 3, the 3PL IRT model was the true model. The 1PL-based CCTs resulted in strikingly lower correct classification error rates of around 82% while the true model based CCTs produced the correct classification rates of around 95%. The differences were very salient. The false positive error rates in the true and the 1PL model based CCTs displayed very sharp contrast. The false positive error rate spanned from 1.9% to 2.82% in the true model based CCT whereas it ranged from 16.38% to 18.3% in the 1PL based CCT that were much higher than the nominal level. The false negative error rates in the 1PL based CCTs lessened conspicuously. The average test length in the true model and the 1PL-based CCTs were alike with one exception.

In Study 3, when the misfit model was 2PL IRT model, there was disparity in the correct classification rates between the true and the misfit model based CCTs. The correct classification rates in the 2PL-based CCTs decreased. The decrease in the false positive error rates in the 2PL-based CCTs was noticeable. On the other hand, the growth in the false negative error rates in the 2PL-based CCTs was remarkable. The false negative error rates in all simulated test conditions were all above 5%. The increase in the average test length in the 2PL-based CCTs was uneven with the largest growth in the test condition with only item exposure control.

[Insert Table 11 about here]

<u>Test Length</u>

In Study 1, when test length constraints was imposed; the correct classification rates went up a little bit, but the differences were not great. The false positive error rates declined. This was also true with the false negative error rates with one exception. Longer average test length was needed for classification decisions when the test length constraints presented.

In Study 2, imposing the test length constraint increased the correct classification rates in most cases. It decreased the false positive and false negative error rates. With test length control, the average test length for classification went up greatly.

In Study 3, when the test length was controlled, the correct classification rates were brought down in most cases. Constraining test length elevated the false positive error rates. The changes in the false negative error rates were not in the same direction when the test length was constrained. Like in Study 2 and Study 3, the average test length used for classification was raised drastically when the test length was controlled.

Item Exposure Control

In Study 1, the item exposure rate control did not affect the percentage of correct classification a lot. It enlarged both the false positive and the false negative error rates very slightly. CCTs with item exposure control generally required longer average test length for classification decisions.

In Study 2, the correct classification rate went down in most test conditions when there was item exposure control. Item exposure control raised the false positive and the false negative error rates in most simulated test conditions. With item exposure control, longer average test length was required for classification decisions.

In Study 3, controlling item exposure rates reduced the correct classification rates. In most cases, item exposure control increased the false positive and the false negative error rates. The average test length used for classification increased when item exposure rate was controlled.

To sum up, when the 1PL or the 2PL IRT model was the true model, the use of a misfit IRT instead of the true model had no serious impact on classification decisions made from CCT. However, when the 3PL IRT model was the true model, the use of the 1PL increased the false positive error rates to extremely intolerable level; and the use of the 2PL raised the false negative error rates to the level higher than the nominal value. In most cases, the use of a misfit IRT model in CCT lowered the correct classification rates and raised the average test length used for classification decisions. The constraint of the test length changed the average test length for classification greatly, but had almost negligible effect on the correct classification rates and error rates. When there was no test length constraint, item exposure control only increased the average test length for classification noticeably. It had almost not serious effect on the correct classification rates and error rates. With test length constraint, item exposure control raised the average test length for classification slightly and almost did not affect correct classification rates and error rates.

## Discussion

When the 1PL and 2PL IRT model was the true model, the impact of using a misfit model was not serious. However, when the 3PL IRT model was the true model, the use of the 1PL was a concern, so is the use of the 2PL. The reasons for the extreme change in the false positive and the false negative error rates in the 3PL based CCT were explored and are explained as follows.

In a true model based CCT using the true item parameters, all items were ranked based on the item information obtained from the true model and the true item parameters. In CCT using a misfit IRT model, true item parameters were employed in item response generation. However, these items were rearranged based on the item information computed from the calibrated item parameters and the model used in calibration. Thus, the items with the true parameters for item response generation in CCT were ranked differently from their original ranking. Different sets of items were selected in CCTs using the true and the calibrated parameters. Furthermore, items with the calibrated parameters were applied in making classification decisions. The likelihood of generating a certain response using the calibrated item parameters and the rearranged true item parameters were not the same. The discrepancy in the probability of correct responses would increase or decrease the two Types of error rates. This can be illustrated by examining the average level of the test difficulty or the average test characteristic curve (ATCC) for a CCT. ATCC is plotted by calculating the average probability of a correct response at each theta point in the range of ability scale –4 to 4. ATCC was obtained by averaging the probability of correct responses for the average number of the items selected in a CCT. Three such curves will be plotted in one figure to demonstrate the differences in the probability of correct responses in the CCT using the true item parameters, the CCT using the rearranged true item parameters for item response generation, and the CCT using the calibrated items parameters for classification decisions. The CCT using the true item parameters is labeled with suffix –TM. The CCT for item response generation has the label with suffix –RA, and the CCT for classification decisions is labeled with suffix –RK. The ATCC for the CCT-TM, the CCT-RA, and the CCT-RK are plotted in Figure 1 to Figure 4 only for the CCT with test length constraint and with no constraint. When the item exposure was controlled, randomness was built in the item selection; thus, such comparison is not meaningful. In all figures, the solid line represents the ACCT for the CCT using the true item parameters. The dashed line stands for the ATCC for the CCT for item response generation. The dotted line is the ATCC for the CCT for classification decisions.

[Insert Figure 1 and Figure 2 about here]

The largest error increase was in the 1PL-based CCT when the 3PL IRT mode was the true model. Figures 1 and 2 show the ATCC for the first 23 and first 63 items. Twenty-three items were the average number of items used for classification decisions in CCT with no constraints. Sixty-three items were the average number of items used for classification decisions in CCT with only test length constraint. Both figures showed a similar pattern. In these two figures, 3PL-TM labels the 3PL-based CCT using the true item parameters ranked based on the information from the true item parameters and the true model. 3PL-RA stands for the 3PL-based CCT with the rearranged true item parameters for item response generation. 1PL-RK is the label for the 1PL-based CCT using the calibrated item parameters for classification decisions. In the ability range of –4 to 4, the 3PL-based CCT for item response generation showed a higher average probability of correct responses than the 1PL-based CCT for classification decisions. This indicated that it was more likely to generate correct responses than the 1PL model should have. Then, the 1PL-based CCT would be more likely to classify examinees as a master. Thus, the false positive error would increase.

[Insert Figure 3 and Figure 4 about here]

When the 2PL IRT model was used instead of the true model, the 3PL IRT model, Figures 3 and 4 display the ATCC for the true model and the 2PL based CCT. In the two ends on the ability scale, the 2PL-based CCT for classification decisions had a higher average probability of correct responses than the 3PL-based CCT for item response generation. Nevertheless, in the middle of the ability distribution, the opposite was true. As the ability was normally distributed, most examinees had the ability in the middle range. Thus, the trend observed in the middle range should have a dominant effect. In the middle range of the ability distribution, the 3PL-based CCT for item response generation was more likely to generate incorrect responses than the 2PL-based CCT for classification decisions should have. Therefore, examinees would be more likely to be classified as a non-master. The false negative error rates increased.

## Conclusions

When the 1PL was the true model, the use of the 2PL or the 3PL IRT in CCT did not lead to serious error rates. However, the 2PL based CCT produced closer results to the 1PL-based CCT than the 3PL IRT based CCT. When the 2PL was the true model, the use of the 1PL and the 3PL IRT model in CCT did not make much difference. When the 3PL IRT was the true model, the use of the 1PL IRT model severely influenced the CCT results. The increase in the Type I error rates could be very striking. The use of the 2PL IRT model as a misfit model affected the CCTs results as well. Under all simulated CCTs, the use of the 3PL IRT model regardless of the true model did not affect the CCT results.

In summary, it is necessary to check the goodness-of-fit of a particular IRT in the implementation of CCT. Without knowing which model fits the data better, it is not valid to stick to one particular IRT model in an IRT model based CCT. Therefore, model evaluation and selection is an indispensable step in IRT based CCT using SPRT.

## Importance

Several studies examined the effect of violation of assumptions of IRT models on classification decisions in CCT. Only Kalohn & Spray (1999) investigated the effect of model misspecification on classification decision based on a computerized test. However, their study was limited to one scenario where the 3-PL IRT model was the true model and the 1-PL IRT model was used in CCT. This study provided a fuller picture of the effect of model misfit on classification decisions based on CCT using SPRT as it investigates more comprehensive scenarios by setting each of the three unidimensional IRT models as the true model.

To choose an IRT model, someone just made the decision based on their own preference regardless of the real nature of test data. They seldom tested the goodness-of-model-fit before using a particular IRT model. The results of this study provided evidence negating such practice in implementing a computerized classification test using sequential probability ratio test.

References

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Issues and Applications*. Boston: Kluwer Nijhoff.

Kalohn J. C., & Spray, J. A. (1999). The effect of model misspecification on classification decisions made using a computerized test. *Journal of Educational Measurement*. 36(1), 47-59.

Lau, C. A. (1996). *Robustness of a unidimensional computerized mastery testing procedure with multidimensional testing data*. Unpublished doctoral dissertation. University of Iowa, Iowa City, IA.

Lau, C. A., & Wang, T. (1998). *Comparing and combining dichotomous and polytomous items with SPRT procedure in computerized classification testing*. Paper presented at the American Educational Research Association Annual Meeting. San Diego, CA.

Lau, C. A., & Wang, T. (1999). *Computerized classification testing under practical constraints with a polytomous model*. Paper presented at the American Educational Research Association Annual Meeting. Montreal, Canada.

Lau, C. A., & Wang, T. (2000). *A new item selection procedure for mixed item type in computerized classification testing*. Paper presented at the American Educational Research Association Annual Meeting. New Orleans, Louisiana.

Lin, C., & Spray, J. (2000). *Effects of item-selection criteria on classification testing with the sequential probability ratio test*. ACT Research Report Series 2000-8. Iowa City, IA: American College Testing.

Parshall, C. G., Spray, J. A., Kalohn, J. C., & Davey, T. (2002). *Practical considerations in computer-based testing.* New York: Springer-Verlag.

Spray, J. A., & Reckase, M. D. (1987). *The effect of item parameter estimation error on decisions made using the sequential probability ratio test.* ACT Research Report Series 87-17. Iowa City, IA: American College Testing.

Spray, J. A., Abdel-fattah, A. A., Huang, C., & Lau, C. A. (1997). *Unidimensional approximations for a computerized test when the item pool and latent space are multidimensional.* ACT Research Report Series 97-5. Iowa City, IA: American College Testing.

Table 1
Research Design

| Sub-study | Simulated CCT Using SPRT | | |
|---|---|---|---|
| | True Item Parameters & True Model | Calibrated Item Parameters | |
| | | Misfit Model | Misfit Model |
| Study 1 | 1PL | 2PL | 3PL |
| Study 2 | 2PL | 1PL | 3PL |
| Study 3 | 3PL | 1PL | 2PL |

Table 2
Descriptive Statistics for the Generated True Parameters

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Person Ability | -3.47 | 3.84 | 4.86E-03 | 0.9978 |
| Item Difficulty | -1.46 | 1.49 | -5.51E-02 | 0.72 |
| *a*-parameter | 0.4 | 1.59 | 0.91 | 0.28 |

Table 3
Descriptive Statistics for the Calibrated Item Parameters using 2PL in Study 1

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.57 | 1.53 | -5.42E-02 | 0.75 |
| *a*-parameter | 0.85 | 1.09 | 0.9666 | 0.046 |

Table 4
Descriptive Statistics for the Calibrated Item Parameters using 3PL in Study 1

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.43 | 1.44 | 0.07 | 0.67 |
| *a*-parameter | 0.88 | 1.33 | 1.11 | 0.081 |
| *c*-parameter | 0.01 | 0.22 | 0.0695 | 0.043 |

Table 5
Descriptive Statistics for the Calibrated Item Parameters using 1PL in Study 2

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.40 | 1.69 | -1.26E-17 | 0.63 |

Table 6
Descriptive Statistics for the Calibrated Item Parameters using 3PL in Study 2

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.43 | 1.52 | 0.1189 | 0.66 |
| *a*-parameter | 0.47 | 2.06 | 1.02 | 0.2875 |
| *c*-parameter | 0.01 | 0.24 | 0.085 | 0.053 |

Table 7
Descriptive Statistics for the Calibrated Item Parameters using 1PL in Study 3

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.10 | 0.98 | 5.64E-18 | 0.46 |

Table 8
Descriptive Statistics for the Calibrated Item Parameters using 2PL in Study 3

| | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|
| Item Difficulty | -1.95 | 1.46 | -0.51 | 0.68 |
| *a*-parameter | 0.28 | 1.44 | 0.68 | 0.24 |

Table 9
CCT comparisons for Study 1

| | Correct Classification Rate | α | β | Mean No. of Items |
|---|---|---|---|---|
| CCT1P.1PTF1 | 95.76 | 2.03 | 2.21 | 19.85 |
| CCT1P.2PRF1 | 95.37 | 2.34 | 2.29 | 20.68 |
| CCT1P.3PRF1 | 94.79 | 3.94 | 1.27 | 18.96 |
| | | | | |
| CCT1P.1PTS1 | 96.17 | 1.63 | 2.2 | 64.10 |
| CCT1P.2PRS1 | 96.43 | 1.72 | 1.85 | 63.02 |
| CCT1P.3PRS1 | 95.54 | 3.63 | 0.83 | 62.58 |
| | | | | |
| CCT1P.1PTF5 | 95.93 | 2.05 | 2.02 | 23.11 |
| CCT1P.2PRF5 | 95.06 | 2.38 | 2.56 | 24.22 |
| CCT1P.3PRF5 | 94.9 | 3.89 | 1.21 | 21.34 |
| | | | | |
| CCT1P.1PTS5 | 95.74 | 1.96 | 2.30 | 63.73 |
| CCT1P.2PRS5 | 95.99 | 1.96 | 2.05 | 63.82 |
| CCT1P.3PRS5 | 95.25 | 3.65 | 1.10 | 63.44 |

Note: The numbers in the first three columns are percentage numbers.

Table 10
CCT comparisons for Study 2

| | Correct Classification Rate | α | β | Mean No. of Items |
|---|---|---|---|---|
| CCT2PSA.2PTF1 | 95.42 | 2.19 | 2.39 | 13.42 |
| CCT2PSA.1PRF1 | 94.86 | 3.75 | 1.39 | 21.53 |
| CCT2PSA.3PRF1 | 94.8 | 3.69 | 1.51 | 12.37 |
| | | | | |
| CCT2PSA.2PTS1 | 96.43 | 1.74 | 1.83 | 62.21 |
| CCT2PSA.1PRS1 | 95.33 | 3.25 | 1.41 | 63.25 |
| CCT2PSA.3PRS1 | 96.28 | 2.99 | 0.73 | 62.13 |
| | | | | |
| CCT2PSA.2PTF5 | 96.03 | 1.83 | 2.14 | 20.88 |
| CCT2PSA.1PRF5 | 94.73 | 3.70 | 1.57 | 24.28 |
| CCT2PSA.3PRF5 | 93.53 | 4.51 | 1.96 | 40.94 |
| | | | | |
| CCT2PSA.2PTS5 | 95.49 | 2.27 | 2.24 | 63.92 |
| CCT2PSA.1PRS5 | 95.21 | 3.27 | 1.52 | 63.84 |
| CCT2PSA.3PRS5 | 95.55 | 3.39 | 1.06 | 63.51 |

Note: The numbers in the first three columns are percentage numbers.

Table 11
CCT comparisons for Study 3

| | Correct Classification Rate | α | β | Mean No. of Items |
|---|---|---|---|---|
| CCT3PSASC.3PTF1 | 95.41 | 2.31 | 2.28 | 22.24 |
| CCT3PSASC.1PRF1 | 83.51 | 16.38 | 0.11 | 23.49 |
| CCT3PSASC.2PRF1 | 93.8 | 0.36 | 5.84 | 28.47 |
| | | | | |
| CCT3PSASC.3PTS1 | 95.8 | 1.90 | 2.30 | 63.83 |
| CCT3PSASC.1PRS1 | 83.11 | 16.89 | 0.00 | 63.72 |
| CCT3PSASC.2PRS1 | 93.66 | 0.52 | 5.82 | 65.07 |
| | | | | |
| CCT3PSASC.3PTF5 | 95.3 | 2.23 | 2.47 | 34.99 |
| CCT3PSASC.1PRF5 | 82.12 | 17.81 | 0.07 | 26.20 |
| CCT3PSASC.2PRF5 | 93.49 | 0.43 | 6.08 | 43.10 |
| | | | | |
| CCT3PSASC.3PTS5 | 94.47 | 2.82 | 2.71 | 66.79 |
| CCT3PSASC.1PRS5 | 81.69 | 18.30 | 0.01 | 64.16 |
| CCT3PSASC.2PRS5 | 92.97 | 0.83 | 6.20 | 68.31 |

Note: The numbers in the first three columns are percentage numbers.

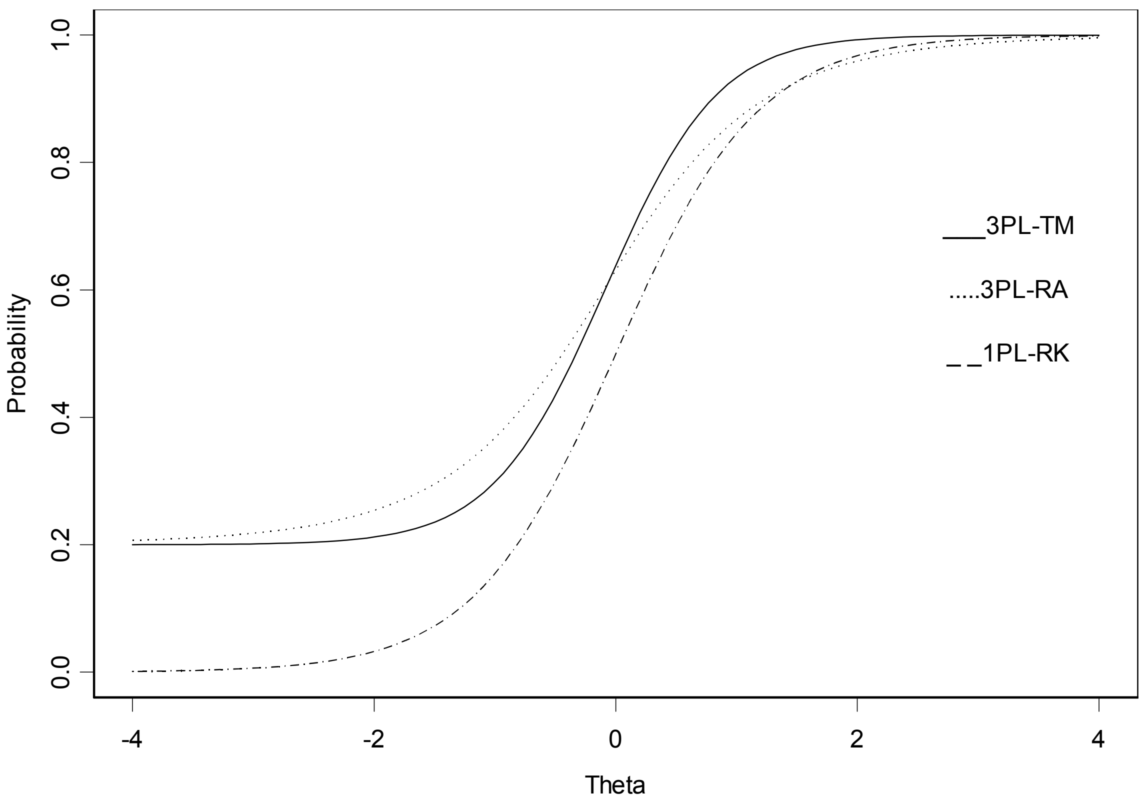


Figure 1. The 3PL-TM and 3PL-1P average characteristic curve for the first 23 items

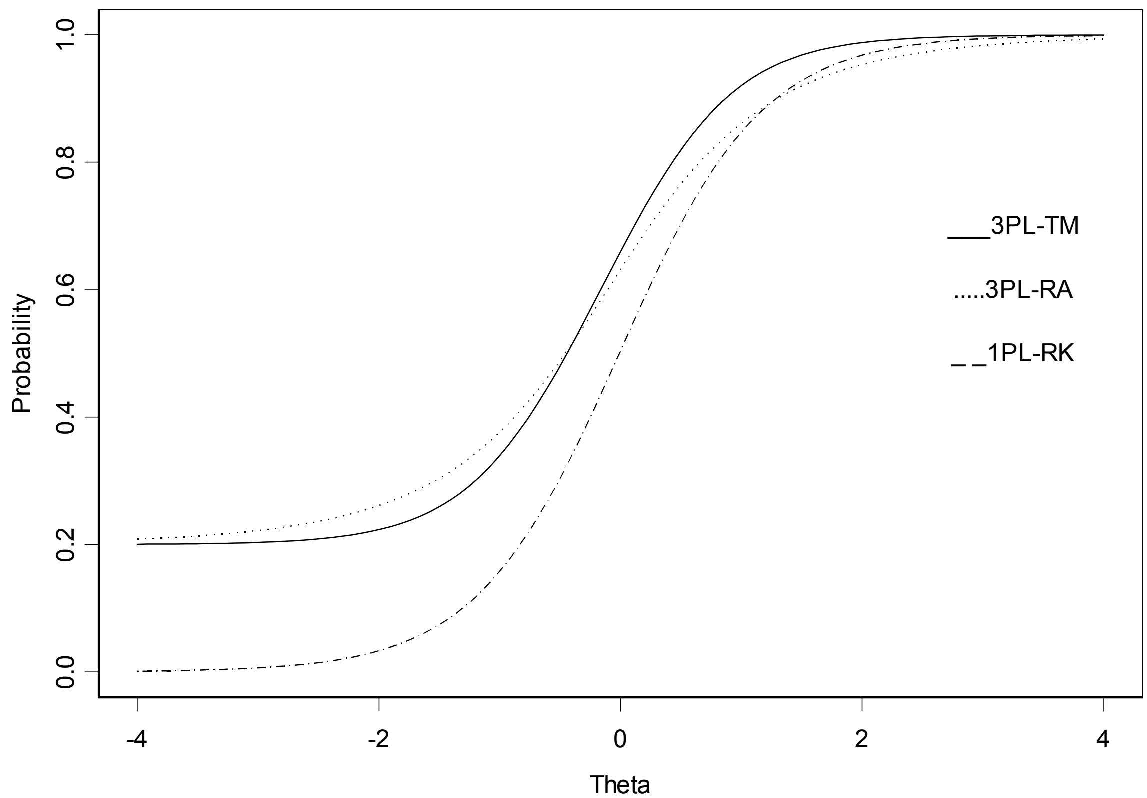


Figure 2. The 3PL-TM and 3PL-1P average characteristic curve for the first 63 items

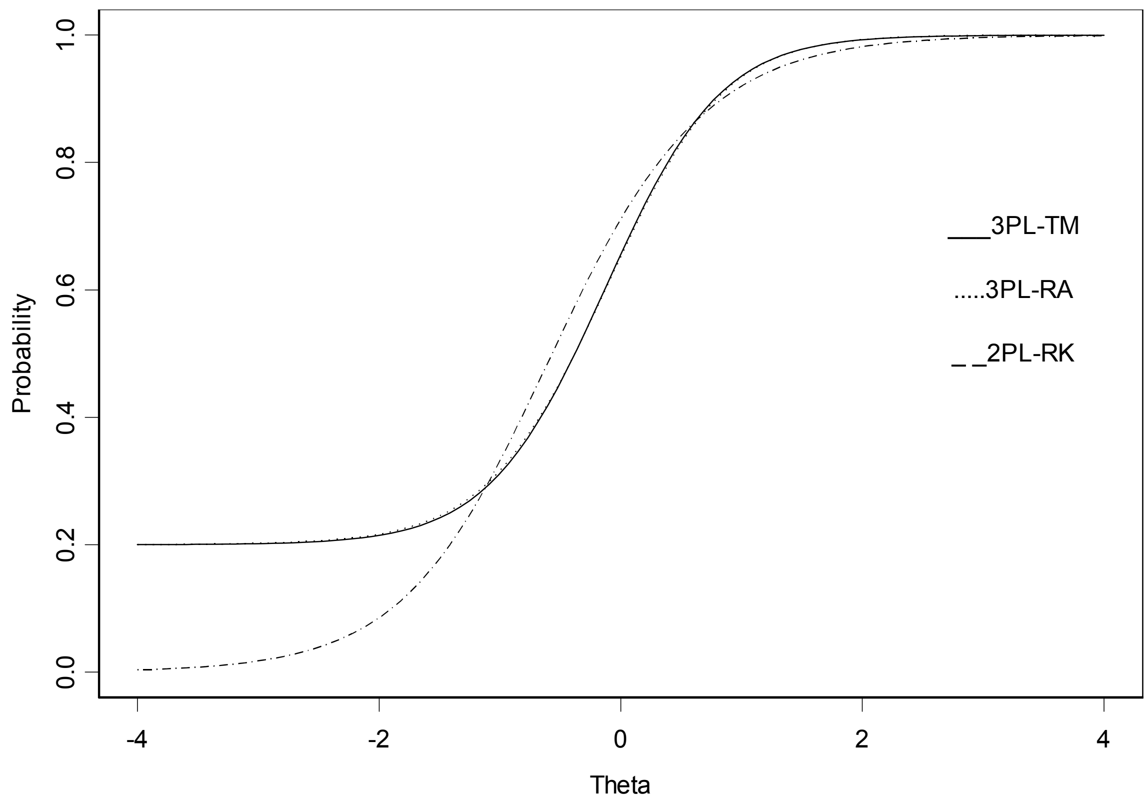


Figure 3. The 3PL-TM and 3PL-2P average characteristic curve for the first 28 items

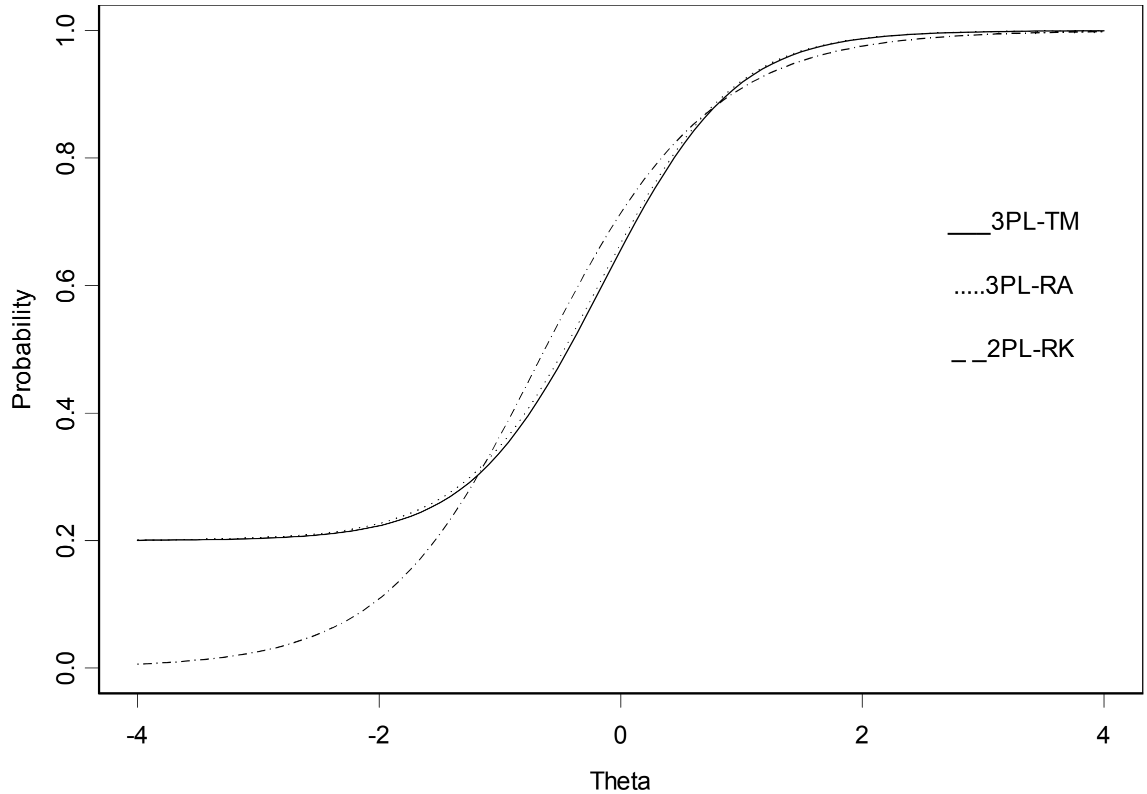


Figure 4. The 3PL-TM and 3PL-2P average characteristic curve for the first 65 items