ISSN: 1309 – 6575
**Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi**
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2022; 13(3); 256-268

# The Effect of Aberrant Responses on Ability Estimation in Computer Adaptive Tests

Sebahat GÖREN* Hakan KARA** Başak ERDEM KARA***
Hülya KELECİOĞLU****

**Abstract**

In computer adaptive test (CAT), aberrant responses caused by some factors such as lucky guesses and carelessness errors may cause significant bias in ability estimation. Correct responses resulting from lucky guesses and false responses resulting from carelessness or anxiety may reveal aberrant responses and the impact of these types of aberrant responses may cause an erroneous estimation of the examinee's actual ability because they do not reflect the examinee's actual knowledge. In this study, the performances of regarding ability estimation were examined comparatively in the context of CAT simulations in case of aberrant responses.Under different conditions, twelve different CAT simulations were conducted with 10 replications for each of the conditions. Correlation, RMSE, bias, and mean absolute error (MAE) values were calculated and interpreted for each condition. Results generally indicated that the 4PL IRT model provided a more efficient and robust ability estimation than the 3PL IRT model and the 4PL model increased the precision and effectiveness of the CAT applications.

*Keywords:* Computer adaptive tests (CAT), 3PL IRT model, 4PL IRT model, aberrant responses, early mistake

## Introduction

Nowadays, many achievement tests, most of which are applied as multiple-choice, are carried out for different purposes such as selection, placement, classification, and evaluation, especially in the field of education. The process of preparing, applying, and evaluating these tests in order to estimate the latent characteristic of individuals in the most appropriate way also changes with the advancement of knowledge and technology. In recent years, the popularity and use of CAT applications, which minimize random errors by providing items appropriate to the individual's ability level, and thus provide the opportunity to reach more accurate information, has increased. So that, individuals only answer items that are appropriate for their ability levels, the length of the test is shortened and the test duration decreases (Thompson, 2009; Wainer, 2000; Weiss, 2004). CAT applications are mostly based on IRT models and IRT models allow for estimation of abilities and comparisons between individuals, even when individuals answer different items at different difficulty levels The test algorithm applied in CAT consists of three basic steps; the starting rule, the progression rule, and the termination rule (Wainer, 2000). Those steps include the application of a predetermined starting rule using an item pool of a sufficient qualified and number of items (starting rule), the selection of the most appropriate item for individual's ability level from the pool based on the temporary ability level calculated after each answered item (progression rule) and the termination of the test based on a specified termination rule (Segall, 2004; Thompson & Weiss, 2011).

Individuals' answers on multiple choice tests are classified into three categories; responses reflecting true ability, correct responses given by chance (lucky guesses), and false responses resulting from anxiety, carelessness, or distraction (Liao et al., 2012). The last two categories, which contain unusual

* Research Assistant, Hacettepe University, Faculty of Education, Ankara-Turkey, sebahatgoren@gmail.com, ORCID ID: 0000-0002-6453-3258
** PhD student., Hacettepe University, Faculty of Education, Ankara-Turkey, hakankaraodtu@gmail.com, ORCID ID: 0000-0002-2396-3462
*** Assist. Prof. Dr., Anadolu University, Faculty of Education, Eskişehir-Turkey, basakerdem@anadolu.edu.tr, ORCID ID: 0000-0003-3066-2892
**** Prof. Dr., Hacettepe University, Faculty of Education, Ankara-Turkey, hulyaebb@hacettepe.edu.tr, ORCID ID: 0000-0002-0741-9934

To cite this article:
Gören, S., Kara, H., Erdem-Kara, B., & Kelecioğlu, H. (2022). The effect of aberrant responses on ability estimation in computer adaptive tests. *Journal of Measurement and Evaluation in Education and Psychology, 13*(3), 256-268. https://doi.org/10.21031/epod.1067307

*Received: 02.02.2022*
*Accepted: 10.07.2022*

answers that we are not used to, do not reflect the true ability level of the individual and cause an erroneous estimate of the individual's true ability. The effect of these abnormal responses is limited due to equal weighting of items in traditional tests based on classical test theory. However, IRT is highly sensitive to that kind of response disturbance since it is a statistical method based on an examinee's response to explain the ability level (Magis, 2014). The existence of aberrant responses may cause a strongly biased estimation of true underlying ability and may jeopardize the accuracy of measurements and invalidate the IRT use (Jia et al., 2019). Since CAT applications are also based on IRT models, they are also open to the same kind of biased estimations and erroneous measurements. Rulison and Loken (2009) stated that chance (luck) factors and attention errors that occur especially at the beginning of a test in CATs may have a significant effect on test results. To solve this problem, Barton and Lord (1981) developed 4PL IRT model from IRT models by adding the inattention (carelessness) parameter ($d_i$) to 3PL IRT model.

With 4PL IRT model, the probability of individuals with high ability levels giving wrong answers for easy items because of the factors such as carelessness, fatigue, or anxiety is calculated. Inattention parameter ($d_j$) allows the upper asymptote to get values smaller than 1.00 and differ between 0.00 and 1.00 theoretically. With the inclusion of an upper asymptote with a value less than 1.00, it is allowed that the place of a high-ability individual does not change significantly in the ability scale in case of a false response to an easy item. Barton and Lord (1981) conducted analysis by fixing upper asymptote values 1.00, 0.99, and 0.98 respectively. In other words, they specified a common upper asymptote value for all items and did not mention the freely estimating d parameter (Waller & Reise, 2010).

With CAT applications, making the temporary ability estimations of each individual after each item and selection of item based on those temporary ability levels makes the use of 4PLM quite meaningful. Since 4PLM aims to estimate the ability of individuals with the high ability with the least error, it is more affected by errors such as carelessness, especially at the beginning of the test, which causes estimation bias (Rulison & Loken, 2009). For example, when the d parameter of the item is in the range of 0-1, when the individual answers this item incorrectly, there will be a decrease in the individual's ability level and this decrease will be less than the other incorrectly answered items. Thus, an item more appropriate for the ability level of the individual can be selected from the item pool later.

Rulison and Loken (2009) made ability estimations with ordinary (not intervened) performance, intervened performances such that individuals intentionally answered the first two items incorrectly and again intentionally answered the first two items correctly in order to investigate the effect of upper asymptote on ability estimation in CAT applications under 3PLM and 4PLM by conducting a simulation study. It was concluded that 4PLM may reduce the estimation error for high-ability examinees who answered the first two items incorrectly. In that study, the analyzes were made by fixing the d parameter to 0.98. Besides, Loken and Rulison (2010) indicated how to make parameter estimates with 4PLM by using both real and simulation data, and at the same time, they compared models by making estimations with 2PLM and 3PLM. While the correlation coefficients were similar for these two data types, the error values were at the lowest level for 4PLM. When estimated item parameters under 4PLM were examined it was observed that d parameter got values between 0.72 and 0.89. On the other hand, Waller and Reise (2010), used Low Self-Esteem Scale of the Minnesota Multiphasic Personality Inventory (MMPI-A) to examine compatibility of data with 4PLM. In that study, ability estimations were made under 3PLM and 4PL models and it resulted that ability estimation did not differ significantly, but only when the relationship between estimations of individuals at high ability level (with low self-esteem) was considered, model selection led to a difference. When the standard errors related to estimations were examined, it was observed that 4PLM had a more accurate estimation with less error. In addition, Liao et al. (2012) conducted a simulation study to compare the measurement precision and efficiency of 3PL and 4PL models under ordinary and poor-start testing conditions. CAT application was carried out under two different conditions; ordinary (not intervened) and the condition in which the first two items were intentionally evaluated as incorrect. Besides, estimations were made according to both 3PL and 4PL models in each of these conditions, and results were compared. When estimation was made with 4PLM, d parameter was fixed to di = 0.98 for all items. It was found that there was no significant difference between models in both ordinary and intervened conditions. When ability estimations were made under 4PL model, the error level was significantly lower than the error on 3PL model.

When individuals take achievement tests in CAT, they can give incorrect answers to items because of some factors such as anxious, careless, and poor testing conditions although they know the correct answer. Those answers are some examples of aberrant responses, and those kinds of answers may cause significant bias in ability estimation especially when they are done at the beginning of the test since estimated thetas do not reflect the examinee's actual knowledge. So, the impact of these types of aberrant responses may cause an erroneous estimation of the examinee's actual ability because they do not reflect the examinee's actual knowledge. To cope with the effect of that aberrant responses, the use of 4PL IRT model is suggested in the literature. It is stated that 4PL IRT model may provide a more efficient and robust ability estimation than the 3PL IRT model in the CAT applications. Rulison and Luken (2009) stated that aberrant responses may influence the CAT results especially if they occur as early mistakes. 4PL IRT model improves the efficiency and precision of CAT under both ordinary conditions and the existence of aberrant responses. It helps to reduce the precision and efficiency degradation caused by careless mistakes (Liao et al., 2012). Despite of that, studies regarding the 4PL IRT model used for aberrant responses on computer adaptive tests (CAT) are limited to a few numbers of studies in the literature (Liao et al., 2012; Rulison & Loken, 2009). In the context of this study, the performances of 3PL and 4PL IRT models on ability estimation were investigated under the conditions in which ordinary (not intervened), the first item was intentionally evaluated as incorrect, and the first two items were intentionally evaluated as incorrect according to different termination rules. The results of the study are likely to contribute to the literature focusing on aberrant responses in computer adaptive test applications.

### Purpose of the Study

In the context of this study, the performance of 3PLM and 4PLM as an error correction mechanism on CAT in case of the existence of aberrant responses was investigated under 12 conditions. In the study conducted for this purpose, an answer to the following research question was sought:

How do the performances of 3PL and 4PL models regarding ability estimation in CAT applications differ according to different response behaviors (ordinary, incorrect answers to the first item and the first two items) and different termination rules (fixed length and varying length)?

## Method

In this study, performances of 3PL and 4PL models regarding ability estimations in CAT applications were examined under different response behaviors and different termination rules generated simulatively.

### Data Generation

In this study, 13 ability levels ranging from -3.0 to +3.0 in equally spaced intervals of .5 were specified and the ability parameters of a total of 2600 individuals, including 200 individuals for each of the 13 levels were generated two different item pools each consisting of 300 items were generated by using 3PL and 4PL IRT models. The a, b, and c parameters of items in the item pool were generated by using log-normal distribution L(0, .25), uniform distribution U[-3, +3], and uniform distribution U[0, .25] respectively. The d parameter in the pool generated using 4PL IRT model was taken as 0.98, based on the findings obtained by Rulison and Loken (2009). After the generation of ability parameters and construction of the item pool, response patterns of examinees were generated and continued with CAT simulation.

## Procedure

In CAT simulation, two different IRT models (3PL and 4PL); two different termination rules (30 items and SE< .30); and three different response behaviors (ordinary, poor-start1, and poor-start2) were used and 12 conditions (***2 IRT models x 3 response behaviors x 2 termination rules = 12 conditions***) were investigated in total. Twelve different CAT applications were applied to each participant for each of the conditions with 10 replication. Individuals' response to the first item in tests with Poor-start1 response behavior and to the first two items in tests with Poor-Start2 response behavior was intentionally evaluated as "incorrect" regardless of their true ability level. Those manipulated items were of medium difficulty. The remaining items were answered according to their true (actual) ability levels. Examinees answered all items according to their true ability levels in ordinary response behavior test conditions. The 12 different conditions are given below:

- O3CAT: Fixed-length CAT based on the 3PL model under ordinary response behavior
- O4CAT: Fixed-length CAT based on the 4PL model under ordinary response behavior
- $O3CAT_s$ : Variable-length CAT based on 3PL model under ordinary response behavior
- $O4CAT_s$ : Variable-length CAT based on 4PL model under ordinary response behavior
- P1_3CAT : Fixed-length CAT based on 3PL model under Poor-Start1 response behavior
- P1_4CAT : Fixed-length CAT based on 4PL model under Poor-Start1 response behavior
- $P1\_3CAT_s$ : Variable-length CAT based on 3PL model under Poor-Start1 response behavior
- $P1\_4CAT_s$ : Variable-length CAT based on 4PL model under Poor-Start1 response behavior
- P2_3CAT: Fixed-length CAT based on 3PL model under Poor-Start2 response behavior
- P2_4CAT: Fixed-length CAT based on 4PL model under Poor-Start2 response behavior
- $P2\_3CAT_s$: Variable-length CAT based on 3PL model under Poor-Start2 response behavior
- $P2\_4CAT_s$: Variable-length CAT based on 4PL model under Poor-Start2 response behavior

For each condition, the ability level for starting rule was specified as '0' and Maximum Fisher Information (MFI) method was used as the item selection method. In order to prevent the same item taking by each individual, randomesque method was used with a 5-item group. The expected a posteriori (EAP) method was preferred for ability estimation and a 0.50 value was used for item exposure.

## Data Analysis

In data analysis, performances of 3PL and 4PL models regarding ability estimations were compared with the help of examining values regarding measurement precision. Those values which were the Pearson correlation coefficient between true ability and estimated ability levels, bias, and RMSE were calculated by taking the average of 10 replications for each condition.

RMSE was calculated by using the following formula. It is the root mean square error on all conditions.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)^2}{n}} \quad (1)$$

Bias is the mean difference between an individual's true ability and estimated ability level as a result of simulation (Miller & Miller, 2004). It is calculated by using the following formula;

$$Bias = \frac{\sum_{i=1}^{n}(\widehat{\theta_i}-\theta_i)}{n} \quad (2)$$

On the other hand, mean absolute error (MAE) is the mean average difference between individuals' estimated ability level and true ability level.

$$MAE = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|}{n} \quad (3)$$

## Results

In the context of this study, performances of 3PL and 4PL models regarding ability estimations were examined under different response behaviors (Ordinary, Poor-Start1, and Poor-Start2) and different termination rules (30 items and SE< .30). Correlation between true and estimated ability levels, RMSE, bias and mean absolute error (MAE) values were calculated and presented in Table 1. Besides, values in Table 1 were visualized and graphs regarding correlation, RMSE, and MAE for each test were presented in Figure 1. Obtained results were interpreted by considering both the table values and figures together.

**Table 1**

*Correlation (r), RMSE, Bias and MAE Values of Tests*

| Response Behaviors | Test | r | RMSE | Bias | MAE |
|---|---|---|---|---|---|
| Ordinary | O3CAT | 0.984 | 0.358 | -0.007 | 0.285 |
| | O4CAT | 0.983 | 0.369 | -0.01 | 0.293 |
| | $O3CAT_S$ | 0.985 | 0.338 | 0.002 | 0.269 |
| | $O4CAT_S$ | 0.986 | 0.335 | 0 | 0.266 |
| Poor - Start1 | P1_3CAT | 0.982 | 0.419 | -0.11 | 0.335 |
| | P1_4CAT | 0.981 | 0.419 | -0.086 | 0.333 |
| | $P1\_3CAT_S$ | 0.985 | 0.412 | -0.14 | 0.333 |
| | $P1\_4CAT_S$ | 0.985 | 0.375 | -0.104 | 0.302 |
| Poor - Start2 | P2_3CAT | 0.982 | 0.435 | -0.132 | 0.347 |
| | P2_4CAT | 0.981 | 0.445 | -0.128 | 0.352 |
| | $P2\_3CAT_S$ | 0.985 | 0.378 | -0.099 | 0.302 |
| | $P2\_4CAT_S$ | 0.985 | 0.377 | -0.102 | 0.299 |

When Table 1 and Figure 1a were examined, it was observed that the correlation between individuals' true ability levels generated before simulation and estimated ability levels because of CAT simulations is similar and high (≥.981) across tests. However, the correlation value was higher for variable-length tests compared to fixed-length tests regardless of response behavior and IRT model. Besides, the 3PL model has a higher correlation value for all answering behaviors although it has quite close values with the 4PL model in fixed-length tests. In the variable-length tests, the 4PL model has a higher correlation in ordinary response behavior, while it is almost equal in other conditions.

When Table 1 and Figure 1b were examined, 4PL is the model with a lower error value for all response behaviors in the variable-length test. In the fixed-length tests, the RMSE values for Poor-Start1 condition were equal for both models while the 3PL model had a lower error value in other conditions.

Bias values in Table 1 revealed that the 4PL model had lower bias values than 3PLM for all conditions regardless of response behaviors and termination rules. Except for ordinary response behavior, negative bias values were obtained in the other 10 conditions. Based on this, it can be said that the true ability of

individuals was underestimated in these tests, but this effect is lower for the 4PL model. In order to calculate and interpret the degree of estimation bias, MAE values were calculated and interpreted.

When Table 1 and Figure 1c were examined together, it was seen that the lowest MAE values were obtained in the 4PL model in all response behaviors for the variable-length test. In the fixed-length test, the 3PL model had a lower MAE value in the ordinary and PoorStart-2 response behaviors, and the 4PL model in the other condition.

Based on all these findings, it can be interpreted that 4PLM offers higher measurement precision for variable-length tests on all response behaviors in general. However, 3PLM gives better results for fixed-length tests. For more detailed interpretations, changes in those values on different ability levels were examined and obtained results were presented in Figure 2 - Figure 7.

As seen in Figure 2, under ordinary response behavior, in fixed length tests RMSE values of O3CAT were generally lower across ability levels. That is, in case of using the 3PLM, obtained values were lower compared to the O4CAT. For variable length tests (O3CATs - O4CATs) 3PL and 4PL models presented similar results. Based on these graphs, it is not appropriate to interpret that one model is superior to another. When graphs were examined in general, it can be shown that RMSE values were lower for variable-length tests compared to fixed-length tests across ability levels. Besides, while the RMSE values of the tests at low ability levels were closer to each other, as the skill level increased, the RMSE values also moved away from each other.

**Figure 1**

*Correlation, RMSE and MAE Values of Tests*

*Figure 1a. Correlation*

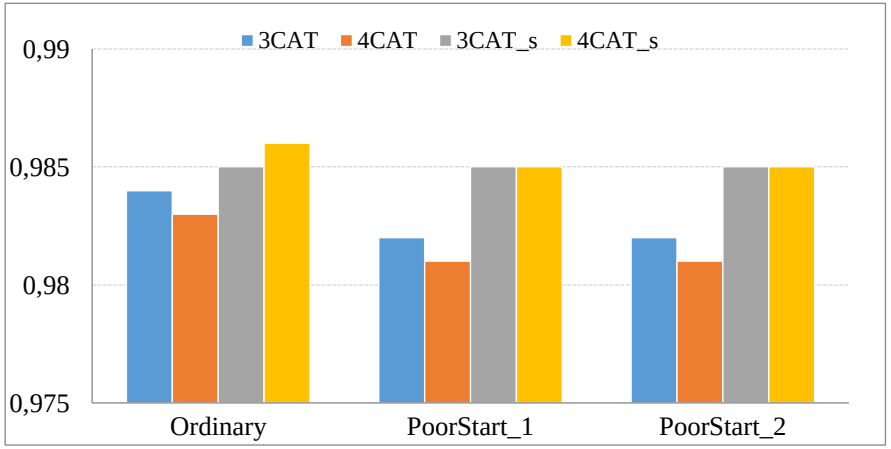



*Figure 1b. RMSE*

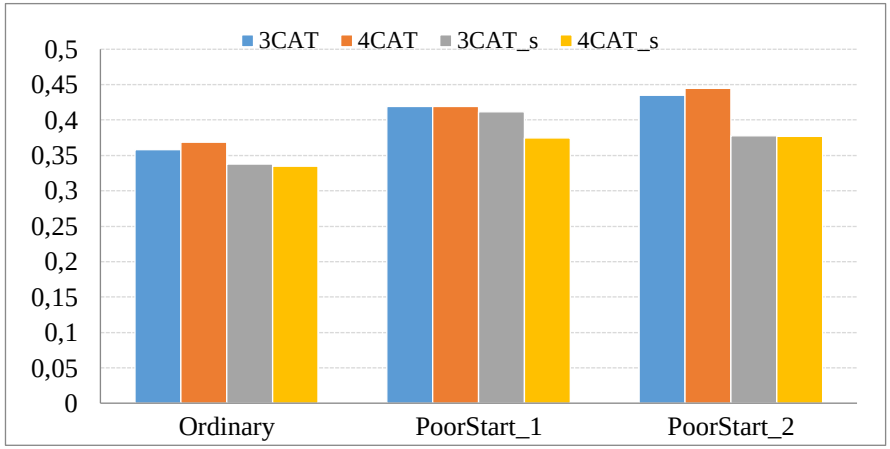



*Figure 1c. Mean Absolute Error (MAE)*

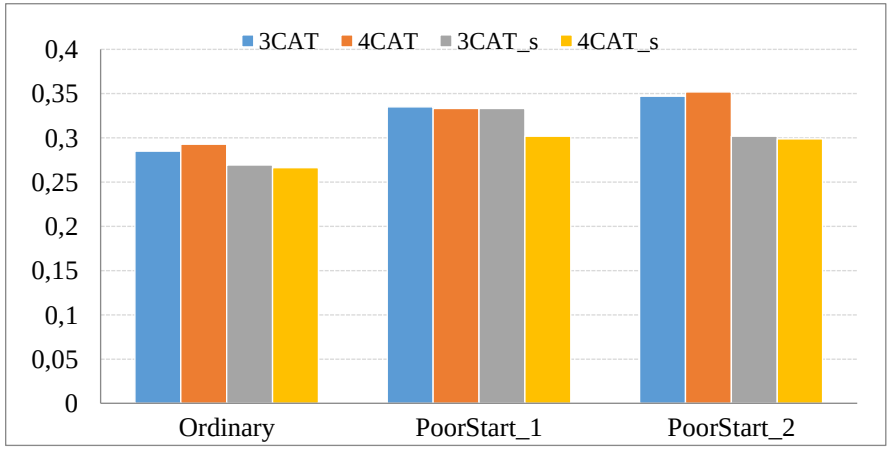
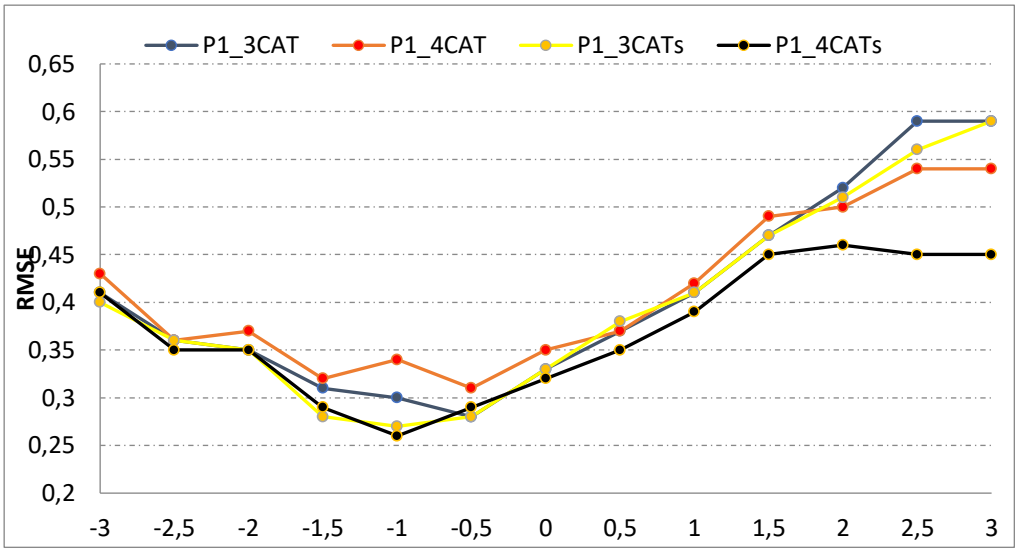


**Figure 3**

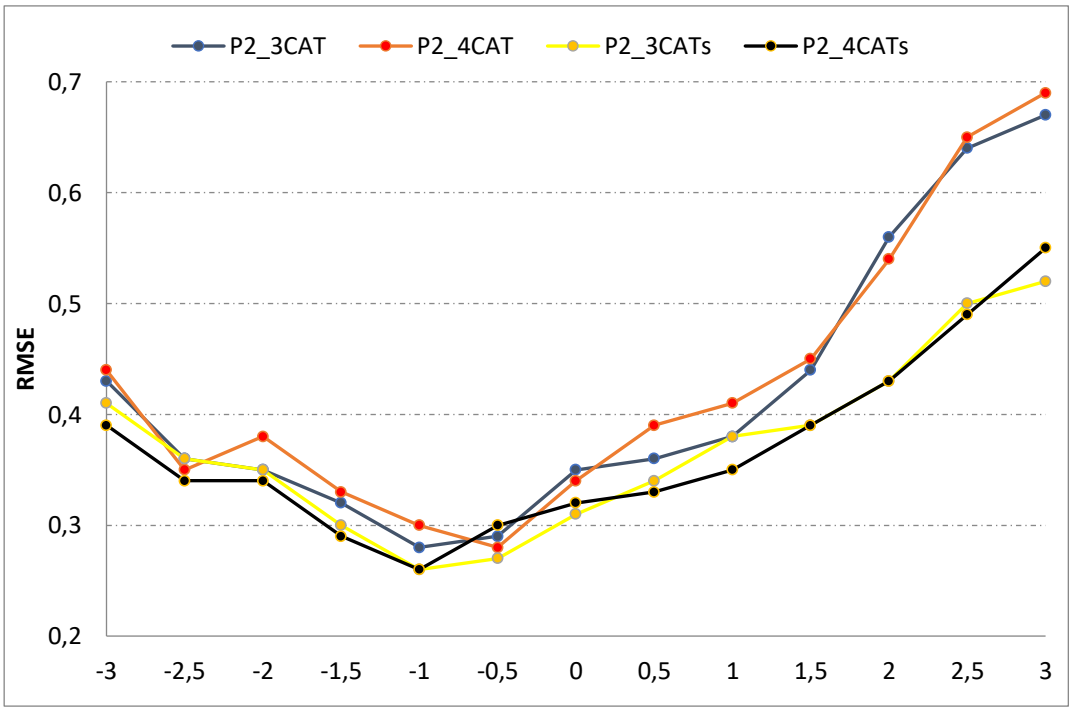
*RMSE Values Across Ability Levels (Poor-Start1)*




As seen in Figure 3, when RMSE values of fixed-length tests (P1_3CAT - P1_4CAT) were compared under Poor-Start1 response behavior; the values of P1_3CAT at low and medium ability levels and the values of P1_4CAT at high ability levels were lower. That is, the ability of individuals with high ability was estimated more accurately with 4PLM. For variable-length tests (P1_3CATs - P1_4CATs), RMSE values of P1_4CATs were lower at high ability levels compared to P1_3CATs and they have similar values at low and medium ability levels. Based on that graph, under Poor-Start1 response behavior, it can be interpreted that 4PLM gave better results. In addition, it was observed that RMSE values of those tests were closer at low and medium ability levels and getting farther at high ability levels.

**Figure 4**

*RMSE Values Across Ability Levels (Poor-Start2)*

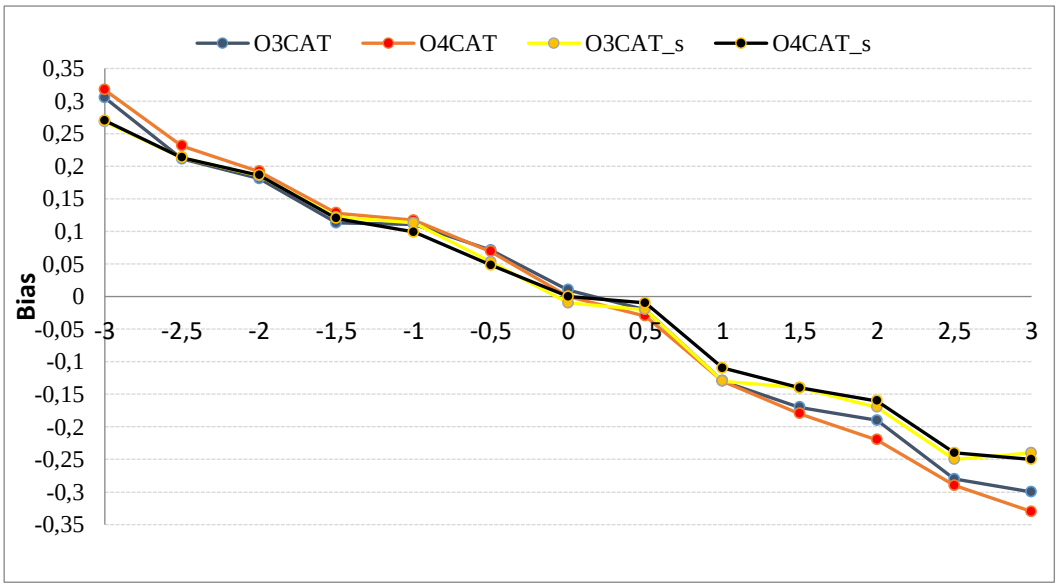


As seen in Figure 4, in terms of RMSE values, both fixed-length tests (P2_3CAT - P2_4CAT) and variable-length tests (P2_3CATs - P2_4CATs) had similar results across ability levels under Poor-Start 2 response behavior. Therefore, it can be said that models do not give superiority to each other. When the graph was examined in general, it can be observed that smaller RMSE values were obtained for variable-length tests compared to fixed-length tests at high ability levels. However, similar RMSE values were found at low and medium ability levels.

As seen in Figure 5, under ordinary response behavior, bias values of tests (O3CAT, O4CAT, O3CATs, and O4CATs) were similar at low and medium ability levels. However, at high ability levels and the lowest ability level ($\theta = -3$), fixed-length tests (N3CAT and N4CAT) were similar to each other, and variable-length tests (O3CATs and O4CATs) were similar to each other in terms of bias values. Therefore, under ordinary response behavior, 3PL and 4PL models were not superior to each other in terms of bias. When the graph was examined in general, the ability of individuals at low ability levels was overestimated (estimated higher than it really is) and the ability of individuals with high ability levels was underestimated (estimated lower than it really is). Besides, at high ability levels, lower bias values were estimated at variable-length tests compared to fixed-length tests, and the ability of individuals was estimated more accurately.

**Figure 5**

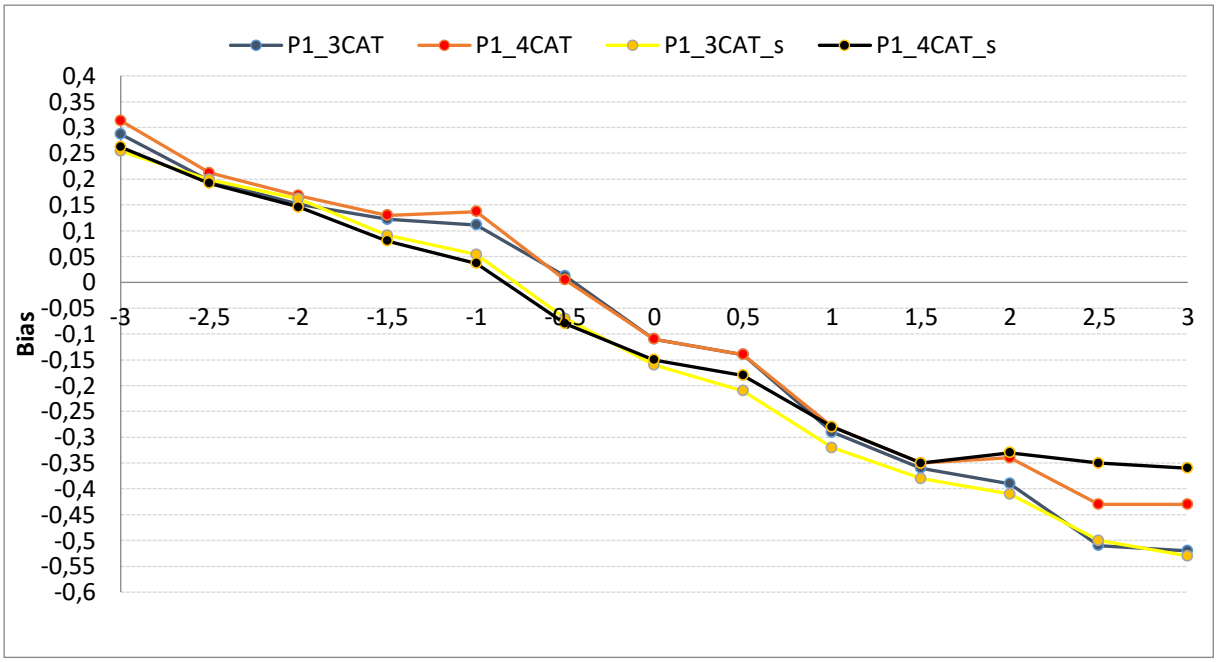
*Bias Values Across Ability Levels (Ordinary)*




According to Figure 6, when the bias values of fixed-length tests (P1_3CAT - P1_4CAT) were compared under the Poor-Start1 response behavior; P1_4CAT values were found to be smaller at higher skill levels, while similar values were calculated at other skill levels. Similarly, variable-length tests (P1_3CATs - P1_4CATs) had lower estimation bias at high ability levels compared to P1_3CAT. Therefore, it can be interpreted that 4PL was better at estimating high-level individuals' ability under Poor-Start1 response behavior.

**Figure 6**

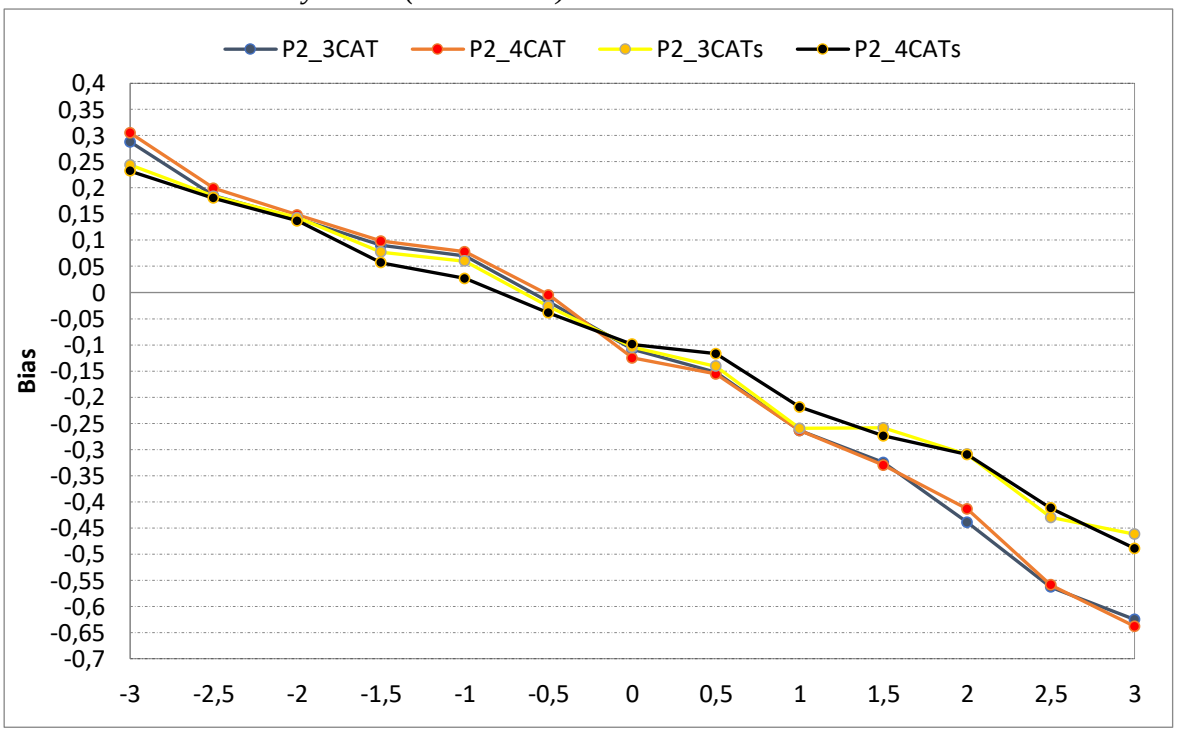
*Bias Values Across Ability Levels (Poor-Start1)*




When Figure 7 was examined, fixed-length tests (P2_3CAT - P2_4CAT) were similar to each other, and variable-length tests (P2_3CATs - P2_4CATs) were similar to each other in terms of bias values at high ability levels under Poor-Start2 response behavior. It was seen that the estimation bias of tests was closer at other ability levels. Accordingly, the models could not provide superiority to each other in the ability estimation of individuals with Poor-Start2 response behavior. In addition, lower bias values were generally estimated in variable-length tests compared to fixed-length tests, and individuals' abilities were estimated more accurately.

**Figure 7**

*Bias Values Across Ability Levels (Poor-Start2)*



## Discussion and Conclusion

In the context of this study, 12 different CAT simulations with different response behaviors (ordinary-not intervened, the intentionally incorrect answer to the first item, and intentionally incorrect answer to the first two items) and different termination rules (30 items and SE<.30) in order to interpret the performances of 3PL and 4PL models regarding ability estimation.

Regardless of response behavior and IRT model, correlation values between individuals' true abilities (generated before simulation) and estimated abilities were higher and RMSE and bias were lower at variable-length tests compared to fixed-length tests. The clear conclusion of this study is that variable-length test performed generally better than fixed-length test in terms of correlation, bias, RMSE, and MAE through all conditions. That finding is consistent with the results of Babcock and Weiss (2012) stating that variable-length CATs performed either slightly better or comparable to the fixed-length test. The reason for the improved efficiency of variable-length tests may be that they allow using of more items which may yield better θ estimates.

In addition, it was concluded that the usage of 4PLM at variable-length tests provided higher measurement precision but 3PLM worked better at fixed-length tests. For the variable-length tests, that is an expected result since it is known that the 4PL model provides for more robust estimations in case of violations of IRT's well-known assumptions (Ackerman, 1989). As a result, aberrant errors that are inconsistent with an examinee's ability affect the ability estimate in 4PL model less than in the 3PL model. Similar to that result, Liao et al. (2012) used variable-length tests (SE<.30) in order to compare models with CAT applications and they concluded that 4PLM gave better results. On the other hand, it is not clear why 3PL model provided higher measurement efficiency for the fixed-length test. Although a certain explanation cannot be made for that, the possible reason may be the test length. Variable-length tests required more items than the fixed ones through all conditions. 4PLM model may have required to use more items to work better.

In addition to those findings, our results also demonstrated that the difference in RMSE and bias between fixed-length and variable-length tests were moving away from each other especially at higher ability levels for each response behavior. This result indicated that the measurement efficiency of fixed and variable-length tests was becoming distant toward high ability levels. In most of the conditions, variable-length tests presented lower RMSE and bias values than fixed-length tests which indicated that variable-length tests had better measurement efficiency for high-ability levels not only in the case of aberrant responses but also the ordinary condition. In addition to that, 4PLM presented either too similar or better results at high ability levels for variable-length tests. The ability of high-level individuals was estimated more accurately with 4PLM and that result is in concordance with the findings in the literature (Liao et al.,2012; Loken & Rulison, 2010; Rulison & Loken, 2009; Waller & Reise, 2010). The reason for the improved efficiency that 4PLM provided for high ability may be that 4PLM is more robust than the 3PLM and the upper asymptote value of 1 in 3PLM failed to accommodate the aberrant responses of high-ability students. On the other hand, the same comments may not be applicable for the fixed-length test. Under ordinary response behavior, 3PLM presented better results compared to 4PLM at estimating the ability of high-ability individuals. Under Poor-Start1 response behavior, it was observed that 4PLM worked better at high level individual's ability estimation. Under Poor-Start2 behavior, it is difficult to say that one model is superior to the other since they behaved in a very similar way. The obtained results are not sufficient to make any explanation or comment for fixed-length tests.

As a result, factors such as being anxious, being affected by poor testing conditions, lack of computer familiarity or misreading the question may cause individuals to give wrong answers to the questions that should have been answered correctly. If those mistakes are made at the beginning of the test, the ability of the individuals can be underestimated under the 3PL model. The results of this study have shown that the 4PL model increases the effectiveness and precision of the CAT in case of aberrant responses, especially for variable-length tests. A more detailed investigation is needed for the fixed-length tests. Additional research should be overtaken to examine this issue under different conditions. Another suggestion for further research is that aberrant responses were handled only as early mistakes in the context of that study. Further research can be made to investigate the effect of aberrant responses which are not an early mistake.

This research does have limitations that could limit the generalizability of our results. The item pools and examinees' ability parameters were specified by the researchers in accordance with the literature and replications were conducted but it is possible to get different results with different item banks and ability distributions.

## Declarations

**Conflict of Interest:** No potential conflict of interest was reported by the authors.

**Ethical Approval:** Data were simulated in the context of this study so, ethical approval is not required.

**Authors Contribution:** Sebahat Gören-Investigation, methodology, software, resources, formal analysis, and writing-original draft. Hakan Kara-Investigation, methodology, visualization, resources, writing. Başak Erdem Kara-Methodology, software, visualization, formal analysis and writing. Hülya Kelecioğlu-Investigation, methodology, supervision, validation.

## References


Ackerman, T. A. (1989). Unidimensional IRT calibration of compensatory and noncompensatory multidimensional items. *Applied Psychological Measurement, 13*(2), 113-127. https://doi.org/10.1177/014662168901300201

Babcock, B., & Weiss, D. J. (2012). Termination criteria in computerized adaptive tests: Do variable-length cats provide efficient and effective measurement? *Journal of Computerized Adaptive Testing, 1*(1), 1-18. https://doi.org/10.7333/1212-0101001

Barton, M. A., & Lord, F. M. (1981). *An upper asymptote for the three-parameter logistic item response model*. (RR 81-20). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.1981.tb01255.x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologist*. Lawrence Erlbaum Associates.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principals and applications*. Kluwer Academic Publishers.

Jia, B., Zhang, X., & Zhu, Z. (2019). A short note on aberrant responses bias in item response theory. *Frontiers in Psychology*, *10*, 43. https://doi.org/10.3389/fpsyg.2019.00043

Liao, W., Ho, R., Yen, Y., & Cheng, H. (2012). The four-parameter logistic item response theory model as a robust method of estimating ability despite aberrant responses. *Social Behavior and Personality*, *40*(10), 1679–1694. https://doi.org/10.2224/sbp.2012.40.10.1679

Loken, E., & Rulison, K. L. (2010). Estimation of a four-parameter item response theory model. *The British Journal of Mathematical and Statistical Psychology*, *63*(3), 509–25. https://doi.org/10.1348/000711009X474502

Magis, D. (2014). On the asymptotic standard error of a class of robust estimators of ability in dichotomous item response models. *The British Journal of Mathematical and Statistical Psychology*, *67*(3), 430–450. https://doi.org/10.1111/bmsp.12027

Miller, I. & Miller, M. (2004). *John E. Freund's mathematical statistics with applications* (7th ed.). Prentice Hall.

Reckase, M., D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. Springer.

Rulison, K. L., & Loken, E. (2009). I've fallen and I can't get up: Can high-ability students recover from early mistakes in CAT? *Applied Psychological Measurement, 33*(2), 83–101. https://doi.org/10.1177/0146621608324023

Segall, D. O. (2004). Computerized adaptive testing. In K. Kempf-Leonard (Ed.), *Encyclopedia of social measurement* (pp. 429-438). Academic.

Thompson, N. A. (2009). Item selection in computerized classification testing. *Educational and Psychological Measurement*, *69*(5), 778-793. https://doi.org/10.1177/0013164408324460

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*(1), 1-9. https://doi.org/10.7275/wqzt-9427

Wainer, H. (Ed.). (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum.

Waller, N. G., & Reise, S. P. (2010). Measuring psychopathology with non-standard item response theory models: Fitting the four-parameter model to the Minnesota Multiphasic Personality Inventory. In S. Embretson (Ed), *New directions in psychological measurement with model-based approaches* (pp. 147-173). American Psychological Association.

Weiss, D. J. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*(2), 71-84. https://doi.org/10.1080/07481756.2004.11909751