ISSN: 1309 – 6575
Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi
Journal of Measurement and Evaluation in Education and Psychology
Research Article; 2023; 14(1); 33-46

The Effects of Different Item Selection Methods on Test Information and Test Efficiency in Computer Adaptive Testing

Merve ŞAHİN KÜRŞAD*

Abstract

The purpose of this study is to examine the effect of different item selection methods on test information function (TIF) and test efficiency in computer adaptive testing (CAT). TIF indicates the quantity of information the test has produced. Test efficiency resembles the amount of information from each item, and more efficient tests are produced from the smallest number of good-quality items. The study was conducted with simulated data, and the constants of the study are sample size, ability parameter distribution, item pool size, model of item response theory (IRT) and distribution of item parameters, ability estimation method, starting rule, item exposure control and stopping rule. The item selection methods, which are the independent variables of this study, are the interval information criterion, efficiency balanced information, matching -b value, Kullback-Leibler information, maximum fisher information, likelihood-weighted information, and random selection. In the comparison of these methods, the best performance in the aspect of TIF is provided by the maximum fisher information method. In terms of test efficiency, the performances of the methods were similar, except for the random selection method, which had the worst performance in terms of both TIF and test efficiency.

Keywords: Computer adaptive testing, test information function, test efficiency

Introduction

In the field of education, where individual differences are important, the use of individually adjusted tests, which are also called “tailored” or “adaptive” tests, has been increasing recently. As computerized adaptive tests (CAT) are individualized tests, each individual encounters different items according to their available ability, and their ability is recalculated after each item application. Therefore, everyone receives different tests (Eggen, 2004; Sulak, 2013).

In CAT applications, each individual is presented with different items according to their estimated ability level. In Item Response Theory (IRT) based CAT applications, individuals’ abilities and item difficulties are placed on the same scale, and their likelihood of answering correctly in the relevant ability is calculated with 50 per cent probability (Lord, 1980). This makes CAT applications preferable in terms of time and cost compared to traditional paper-pencil tests, as CAT applications conclude the test with fewer items and allow for as valid and reliable tests as paper-pencil test applications (Çıkrıkçı-Demirtaşlı, 1999; Kaptan, 1993; Wainer, 1993; Weiss & Kingsbury, 1984).

Item response theory and its models are important factors for CAT applications because, to match and evaluate the item and ability parameters, IRT-based estimations are needed in CAT applications (Thompson & Weiss, 2011). There are different IRT models, such as one parameter logistic model (1PLM), two parameters logistic model (2PLM), three parameters logistic model (3PLM), and these kinds of models are used when item responses are evaluated dichotomously (correct/incorrect or yes/no). Some other models are used when item responses are evaluated polytomously (Brown, 2018), and some of these models are the Partial Credit Model, Generalized Partial Credit Model, Graded Response Model, and Nominal Response Model (Doğan & Aybek, 2021). For dichotomously scored item responses in CAT applications, 3PLM is the most accepted model (Green et al., 1984; Lord,1980; Weiss, 1983; as cited in. Hambleton et al., 1991). Even individuals encounter different items in CAT

* PhD., TED University, Faculty of Education, Ankara-Türkiye, merve.kursad@tedu.edu.tr, ORCID ID: 0000-0002-6591-0705

To cite this article:
Şahin-Kürşad, M. (2023). The effects of different item selection methods on test information and test efficiency in computer adaptive testing. *Journal of Measurement and Evaluation in Education and Psychology, 14*(1), 33-46. https://doi.org/10.21031/epod.1140757

Received: 5.07.2022
Accepted: 25.12.2022

applications, IRT models provide standards to estimate different individuals' abilities (Hambleton et al., 1991).

Computer adaptive tests are implemented on individuals using various algorithms. These algorithms are implemented in three stages that are the start, continuation, and termination stages. CAT applications start with an item selected from the item pool. The individuals' ability is estimated depending on whether they answer correctly, and another item is selected from the item pool. Ability is estimated after each time the individual receives another item. This process continues until a certain level of accuracy is reached in terms of ability estimation if the variable-length test termination rule is selected, and the process stops when a defined number of items are reached if the fixed-length test termination rule is selected. Then the test is concluded. (Eggen, 2004).

A well-defined item pool with high-quality items is necessary for CAT applications to obtain reliable and valid results, as the results are reliable and valid to the extent of the quality of items. In relation to the item pool, the item selection methods are one of the important factors that affect ability estimation, test information and test efficiency (Han, 2009). Item selection methods have been developed to make the use of the item pool more efficient (Boyd, 2003). These methods influence the start, continuation, and termination of the application of CAT because these methods determine which items the individuals should encounter according to their ability level. Aside from the CAT process, another factor that increases the importance of item selection methods is test security. These methods affect test security by providing balance in item pool usage (Sulak, 2013). There are various item selection methods used in CAT applications. Some of these are Interval Information Criterion (IIC), Efficiency Balanced Information (EBI), Matching-b Value (MbV), Kullback-Leibler Information (KLI), Maximum Fisher Information (MFI), Likelihood Weighted Information (LWI), Random Selection (RS), a-Stratification and Gradual Maximum Information Ratio (GMIR) (Han, 2012). These methods are better or stronger than each other in their various aspects, and there are some studies about the comparison of performances of these methods.

Wen et al. (2000) simulation study on the efficiency of MFI and a-stratification with 5,000 individuals and an item pool of 400 items found that the a-stratification method leads to better results compared to the MFI method in terms of efficient use of the item pool. Costa et al. (2009) compared the performance of MFI, KLI and Maximum Expected Information (MEI) methods in the CAT application. The researchers created an item pool of 246 items from the University of Brasilia's English Language Test and performed a simulation study. They used the bias and mean squared error factors while comparing these methods. It was found that all three methods produced lower bias and standard error values as the test length increased.

In Han's (2009) study examining the item selection methods' influence on item use frequency, test information, item pool index, bias and errors of Ө ability estimation in CAT environment, randomly selected MFI, fade-away MFI, GMIR and fade-away GMIR methods were used. According to the simulation study with 250 individuals and 500 items, it was found that although the MFI method was frequently preferred, the GMIR method produced lower standard errors in terms of the variables specified in CAT applications. Han (2010) compared the a-stratification, IIC, LWI, KLI and GMIR methods in the CAT application. The study tried to identify the item selection method which provided the most efficient use of the item pool and the best performance balance in evaluating the performance of the methods. The study was performed as a simulation study and used 500 multiple-choice items from the Graduate Management Admission Test (GMAT) and the number of individuals was simulated as 80,000, 40,000 and 20,000. According to the results of the study, the MFI, KLI and GMIR methods performed better in situations where the test length is relatively shorter.

Sulak (2013) compared the simulative performance of the MFI, a- stratification, LWI, GMIR and KLI methods. The study was simulated with 2,000 individuals and 250 items and examined the item use frequency of the item selection methods. It was concluded at the end of the study that with regards to item pool use, all item selection methods chose items with high discrimination with greater incidence and therefore were not well balanced in terms of item pool use.

Boztunç Öztürk and Doğan (2015) used a- stratification, MFI and GMIR methods in their study, where they compared different item exposure control methods. They made comparisons regarding estimation precision and test security in the study. According to one of the results of the study, MFI and GMIR methods yielded similar results in terms of item pool use, while the a-stratification method balanced item pool use even in situations where the item pool is not controlled and reduced the test overlap ratio.

There were few studies that examined the effect of item selection methods on TIF, and these studies didn't directly examine the TIF, but standard errors, which are the function of TIF, were calculated (Costa et al., 2009; Deng et al., 2010; Han, 2009; Han, 2010; Sulak, 2013; Sulak & Kelecioğlu, 2019). Also, CAT conditions (start, continuation, and termination stages), sample sizes, and item pool sizes in these studies were different from each other and this study. As mentioned above, IRT models are useful for CAT applications. One of its advantages for CAT is the estimation of individual standard error (Sulak, 2013). The standard error is used to calculate TIF, and TIF resembles how well the test is in estimating the ability and how well the test differentiates the individuals at related ability levels. In the TIF estimation process, item information is summed at related ability levels, and so, selected item information by item selection methods can affect the TIF (Baker, 1986; Hambleton et al., 1991). Therefore, it is important to investigate the effect of item selection methods on TIF.

Besides TIF, the item pool is important for CAT applications. Developing a qualitative item pool is a difficult process, and it is essential to use these items in CAT. For test security, items in the item pool should be used equally (Davis, 2002). One of the factors which affect this process is the item selection methods. These methods are important because selecting the item which maximizes the TIF about individuals taking the test is of critical importance to get effective ability estimations (Sulak, 2013). To evaluate the performance of item selection methods in the aspect of test efficiency, generally average test length (ATL), average exposure rate of items (AERI), scaled chi-square value (X^2), underexposed item rate (UIR) and overexposed item rate (OIR) values are evaluated articles in the accessible literature (Boztunç Öztürk & Doğan, 2015; Boztunç Öztürk, 2014; Han, 2009; Lee & Dodd, 2012; Moyer et al., 2012). ATL represents the number of items implemented to each individual; AERI represents the ratio of the total exposure rate of an item to the number of individuals, scaled chi-square value (X^2) represents the difference between the observed and expected item use frequency. This value gives information about how effectively the item pool is used (Chang & Ying, 1999). Underexposed Item Rate (UIR) represents the average exposure rate of items at lower than 0.02; that is, some items are rarely used, and UIR gives the rates of rarely used items. Overexposed Item Rate (OIR) represents the average exposure rate of items at higher than 0.20; that is, some items are used so frequently in CAT, and OIR gives the rate of frequently used items (Eggen, 2001). Investigation of the effect of item selection methods on these factors is vital because item selection methods can affect the usage of items that is balanced in the item pool with which item is applied next, and so it can affect the test security (Sulak, 2013).

A general evaluation of the studies mentioned above, and the literature indicates that the performance of item selection methods varies under different conditions in CAT applications. Generally, item use frequency, test length, bias and root mean square error (RMSE) values were taken into account when comparing item selection methods. Aside from this, the methods examined were the frequently used a-stratification, MFI and KLI methods (Balta & Uçar, 2022; Costa et al., 2009; Sulak, 2013; Wen et al., 2000). The purpose of this study is to compare the effects of the IIC, EBI, matching -b value, LWI and RS methods besides the frequently used a-stratification, MFI and KLI methods on test efficiency and TIF. These item selection methods were chosen because, firstly, a- stratification, MFI and KLI methods were frequently used methods in CAT applications and these methods' effect on TIF and test efficiency is important. Also, IIC, EBI, matching -b value, LWI and RS methods were selected because their item selection algorithms are different from each other and their effect on TIF and test efficiency were not investigated directly in the available literature. Besides these, research in related literature has shown that item selection methods have some advantages and disadvantages according to each other, but in these researches, two item selection methods were compared (Balta & Uçar, 2022; Wen et al., 2001; Yi & Chang, 2003) or five item selection methods were compared via only average test length in the aspect of test efficiency (Sulak & Kelecioğlu, 2019). Also, these studies' CAT application rules (start, continuation and termination stages) were different from each other, and they evaluated test efficiency generally via only average test length. In this study, especially effect of item selection methods on test efficiency factors is investigated more detail. Accordingly, the research questions of this study are:

1. How do TIF values change according to IIC, EBI, MbV, KLI, MFI, LWI and RS item selection methods?
2. How do the average test length (ATL), average exposure rate of items (AERI), scaled chi-square value (X^2), underexposed item rate (UIR) and overexposed item rate (OIR) change according to IIC, EBI, MbV, KLI, MFI, LWI and RS item selection methods?

Method

This section provides information about the research model and data production and analysis.

Research Model

This research is a simulation study carried out with simulated data. Simulation studies are conducted under conditions where there is no real data or when situations that are more complex than real life are examined (Ranganathan & Foster, 2003). Due to the application difficulty of CAT applications and the requirement of a large item pool and large sample size, most studies on CAT are carried out using simulated data (Babcock & Albano, 2012; Deng, Ansley & Chang, 2010; Han, 2010; Sulak, 2013). As this study is also based on IRT-based CAT applications and requires a large item pool, a large sample size and is difficult to apply, it was conducted as a simulation study.

Data Production and Analysis

Production and analysis of research data were carried out through the SimulCAT (Version 1.2) program developed by Han (2012). Large-scale tests and features of CAT applications were taken into consideration in data production. Ability parameters were produced first, followed by item parameters. Constants and independent variables taken into account in simulated data production are shown in Table 1.

Table 1

Constants and Independent Variables Taken Into Account in Simulated Data Production in CAT

Constant Conditions		Independent Variables	
1.	Sample size	**1.**	Item selection method
2.	Distribution of Ability Parameters	**a.**	Interval Information Criterion (IIC)
3.	Item pool size	**b.**	Efficiency Balanced Information (EBI)
4.	IRT model and Distribution of Item Parameters	**c.**	Matching –b Value
5.	Ability estimation method	**d.**	Kullback-Leibler Information (KLI)
6.	Starting rule	**e.**	Maximum Fisher Information (MFI)
7.	Item use frequency control method	**f.**	Likelihood Weighted Information (MFI)
8.	Stopping rule	**g.**	Random Selection

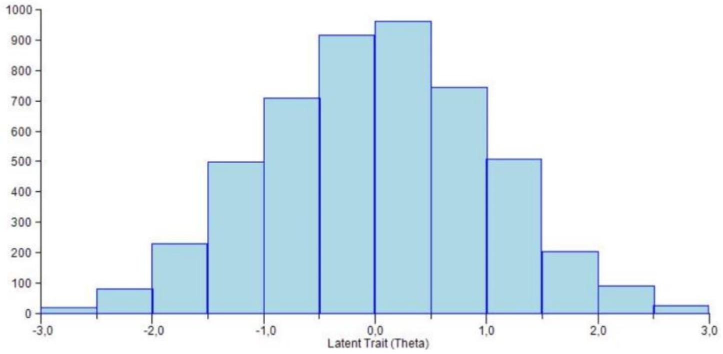
Constants of the Study Taken into Account in Simulated Data Production

Sample Size: It has been claimed that a sample size of at least 1,000 individuals is necessary to accurately perform parameter estimations in IRT-based CAT applications (Rudner & Guo, 2011; Stahl & Muckle, 2007). Besides this, in the study conducted by Şahin (2012), where different IRT models (1PLM, 2PLM, 3PLM) were compared, it was stated that low standard error values were obtained in a sample size of 5,000. Therefore, this study is based on a situation where the sample size is 5,000.

Distribution of Ability Parameters: Ability parameters were produced through a normal distribution with a mean of zero and a standard deviation of one. The reason for choosing normal distribution is that the Expected A Posteriori (EAP) method was selected as the ability estimation method. When the EAP method is used, ability parameters need to be chosen from a universe whose distribution is known, or they need to display normal distribution (Gershon, 2005). The distribution of ability parameters in a sample of 5,000 is given in Figure 1.

Figure 1

Distribution of Ability Parameters in a Sample of 5,000



Item Pool Size: Although it has been stated that an item pool of at least 100 items is necessary to make accurate ability estimations in CAT applications (Urry, 1977), an item pool with such a low number of items is not enough (Sulak, 2013). Stocking (1992) states that the item pool should be at least more than six to ten times the length of the test. In this study, the size of the item pool was set at 500. Because as stated by Risk (2010), an item pool of 500 items is taken to be the ideal item pool size for item pool certificate programs.

IRT model and Distribution of Item Parameters: As the item pool produced for this study consists of dichotomously scored (1/0 answer pattern) items, 3PLM was selected as the IRT model. According to 3PLM, the probability of answering an item correctly is calculated as follows (Birnbaum, 1968);

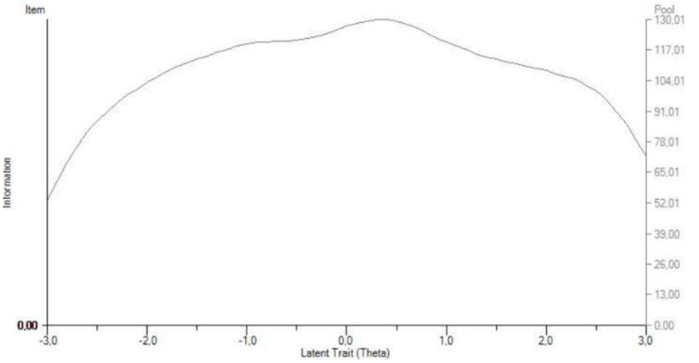
$$P\ (x_j = \mid \Theta_k,\ a_j,\ b_j,\ c_j) = cj + \frac{1-cj}{1+\ e^{-Daj\,(\Theta k - bj)}} \qquad (1)$$

In the formula 1, xj represents the answer to item j (one if true, zero if false), aj represents the slope related to item j, i.e. discrimination, bj represents the threshold parameter for item j, i.e. difficulty, cj represents the low asymptote of item j, i.e. the chance parameter; while the D value represents the scaling constant and has the 1.7 constant value in normal ogive models.

Meanwhile, the distribution of item parameters can vary according to the purpose of the test. According to Boyd (2003), items in the item pool should contain various items ranging from easy to difficult for achievement tests. The ideal distribution of item parameters should consist of a uniform distribution. Therefore, as item production was carried out taking into account achievement tests and the 3PLM was utilized within the scope of this study, the -a parameter was produced from a uniform distribution with the value of between 0.50-2.00; the -b parameter was produced from a uniform distribution with the value of between -3.00-3.00; and the -c parameter was produced from a uniform distribution with the value of between 0.05-0.20. The TIF for the item pool of 500 items according to the specified item parameter values is shown in Figure 2.

Figure 2

The Test Information Function for the Item Pool of 500 Items



Ability Estimation Method: There are many ability estimation methods that are used in CAT applications, such as Weighted Likelihood Estimation (Warm, 1989), Minimum Chi Quadrant (Zwinderman & van den Wollenberg, 1990) etc., two of the most frequently used ones are Maximum Likelihood Estimation (MLE) (Baker, 1992) and Bayesian estimation methods. Bayesian estimation methods come in two types, the Expected A Posteriori (EAP) (Bock & Aitkin, 1981; Bock & Mislevy, 1982) and the Maximum A Posteriori (MAP) (Samejima, 1969). In most of the studies where these methods are compared to each other, it was found that the EAP ability estimation method produced a lower standard error and lower bias values compared to MLE and MAP methods and performed better (Eroğlu, 2013; Keller, 2000; Kezer, 2013; Kingsbury & Zara, 1989). Aside from these studies, Sulak's (2013) study, it was found that when the EAP method is utilized, different item selection methods' standard values produced lower standard errors compared to other ability estimation methods. Therefore, the EAP ability estimation method was used in the current study.

Starting rule: The starting rule covers the selection process of the first items to be selected at the beginning of the test. There are various starting rules in this context. One of them is assigning a value that corresponds to the average ability level. In CAT applications, this value is generally accepted as zero, meaning they start with average items that address zero ability levels. However, when this rule is employed, all individuals receive the same first item, and this can be a problem in the aspect of test security or item exposure. Therefore, random assignment of items can be needed. To eliminate this situation, the starting rule can choose from items with a value of between -0.5 and 0.5. With this method, it is assumed that no information is available on examinees and only several possible items, still not too much, can be assigned to individuals to estimate their initial ability (Thompson & Weiss, 2011). Therefore, this study employs a starting rule that uses items with a value of between -0.5 and 0.5, which corresponds to the average ability level.

Item use frequency control method: Item use frequency methods are used to prevent the repeated use of a certain item. The goal here is to protect the security of the item pool (Davis & Dodd, 2005). There are three items that use frequency control methods named the randomesque method, the Sympson-Hetter method and the fade-away method. In Randomesque method, items are selected from a group of items from the most informative items at the current ability level. However, this method is less successful for dichotomously scored items (Han, 2011; Kingsbury & Zara, 1989; Lee & Dodd, 2012). Sympson-Hetter method gives the conditional probability of an item which will be applied (Boztunç Öztürk, 2014; Han, 2011). Despite the Sympson-Hetter method is well in item pool usage, it is stated that it has a laborious process (Boztunç Öztürk & Doğan, 2015). In the fade-away method, less used and frequently used items' usage are balanced that is frequently used items are used less (Han, 2011). The fade-away method was used in this study as it provides more accurate results compared to other methods in terms of reducing the skewness and test conflict with regard to the item pool usage (Boztunç Öztürk & Doğan, 2015; Davis & Dodd, 2005; Han, 2012).

Stopping rule: The stopping rule refers to the test's termination criterion. There are two stopping rules that are fixed-length and variable-length stopping rules. Studies indicate that the variable-length stopping rule provides a better measurement quality compared to the fixed-length stopping rule and terminates the test more economically (Babcock & Weiss, 2012; Eroğlu, 2013). Used with the stopping rule based on the number of items, it is stated that the distribution approaches normal when 13 or more items are used (Blais & Raiche, 2010). Aside from this, it is stated that better results are obtained in terms of measurement accuracy when the standard error is equal to or lower than 0.40 in the [-3.00; +3.00] ability level range (Babcock & Weiss, 2012; Eroğlu, 2013). Therefore, this study takes into account a situation where the minimum number of items is 15, and the standard error is lower than 0.40 for the test termination rule.

Independent Variables in Simulated Data Production

Item selection methods were taken into account as an independent variables in simulated data production. The explanation of some of these methods, according to Han (2012), is given below:

Maximum Fisher Information (MFI): The MFI method is the most frequently used method and tries to find the item that maximizes the $I_i[\widehat{\theta}_{m-1}]$ value for the m-1 item applied until that instance after each answer given to the items. It selects items through the "local information" around the relevant Θ (Weiss, 1982).

Interval Information Criterion (IIC): The IIC method was developed as an alternative to the MFI. In this method, developed by Veerkamp and Berger (1997), the information function is centered along the confidence interval of the interim Θ estimation. The information functions are averaged along the confidence interval.

Kullback-Leibler Information (KL, Kullback-Leibler Information (Global Information)): The KL method, developed by Chang and Ying (1999), is a function of two ability variables (Θ and Θ_0) and expresses the change of an item between two Θ levels and uses "global information" in item selection.

Likelihood Weighted Information Criterion (LWI): In this method, the information function is collected along the Θ scale and weighted with the probability function.

While IIC, KL and LWI methods perform evaluations along a Θ range using the item information functions, the MFI method selects items based on its evaluation according to the Θ value at a certain point.

Randomization: In this method, items are chosen randomly. As items are selected randomly, this method does not have adaptive testing features (Han, 2012). Despite of its non-adaptive feature, this method was used in research (Choi & Swartz, 2009; Eggen, 2012). For instance, Choi and Swartz (2009) compare the item selection methods as; methods which select the next item randomly and methods which select the next item based on MFI. Eggen (2012) stated that when it was used as an item selection method, efficiency loss was huge, and it affected root mean square error value negatively, especially when selecting harder items. It was stated that random selection or selection of harder items might be motivating in learning systems, which negatively affects ability estimations (Eggen, 2012). Also, the random selection method showed more bias, especially when theta values were away from the mean (Choi & Swartz, 2009). In these studies, the effect of random selection on TIF and test efficiency were not studied. Therefore, in this study, it is used to see its effect on TIF and test efficiency, also.

a-Stratification: The aim of this method is to prevent the selection of high-discrimination items. In this method, developed by Chang and Ying (1999), the items in the item pool are stratified based on their item discrimination, and the item with the -b parameter value close to the interim Θ value is selected from these strata starting with the lowest a parameter value.

Best Matching b-Value (MbV): This method is a special application of the a-stratification method. This method uses only one item stratum and the -b parameter closest to the interim Θ value is selected regardless of the -a parameter and -c parameter.

Gradual Maximum Information Ratio (GMIR): The GMIR method developed by Han (2009) is a method based on expected item efficiency. This value is defined as the effectuation level of information belonging to the item that is closest to the interim Θ.

Efficiency Balanced Information (EBI): In the EBI method developed by Han (2012), unlike the GMIR method, item information is evaluated throughout the Θ interval instead of a certain Θ point. The width of the Θ interval is determined with the standard error of estimation.

The methods included in this study, among those listed above, are the IIC, EBI, MbV, KLI, MFI, LWI and RS methods. GMIR and a-stratification methods were excluded from the study as they failed to initiate the CAT process under specified stopping rules and item pool characteristics.

A total of seven different item selection methods were compared in simulated data production. Twenty-five replications were performed for each item selection method, and 175 analyses were conducted.

The ATL, AERI, X^2, UIR, OIR were calculated while determining test efficiency and also TIF values were calculated for comparing the seven-item selection method during the data analysis process, and these can be named as dependent variables of this study. Calculations were performed through standard errors in the TIF. The formulas regarding the data analysis process are given in Table 2.

Table 2
Criteria for Evaluating Test Efficiency and Test Information Function

	Criteria	Description	Formula
Test Information Function	**TIF**	Equal to the information function of the items in the test taken by individuals in the CAT application	Sem($\hat{\Theta}$)=$\frac{1}{\sqrt{I(\theta)}}$
Test Efficiency	**Average Test Length (ATL)**	The number of items implemented for each individual	$\sum_{i=1}^{n}\frac{Ki}{n}$
	Average exposure rate of items (AERI)	The ratio of the total exposure rate of an item to the number of individuals	$\sum_{i=1}^{n}\frac{mk*100}{n}$
	Scaled chi-square (X^2**)**	Difference between the observed and expected item use frequency	$\sum_{I=1}^{N}\frac{\left(Koj-\underline{K}oj\right)^2}{\underline{K}oj}$
	Underexposed item rate (UIR < 0.02)	Average exposure rate of items at lower than 0.02	$\sum_{i=1}^{n}\frac{mk*100}{n}$< 0.02
	Overexposed item rate (OIR > 0.20)	Average exposure rate of items at higher than 0.20	$\sum_{i=1}^{n}\frac{mk*100}{n}$> 0.20

n: total number of individuals, *Ki*: i. the number of items implemented to each individual, *mk*: the number of times item k is applied to all individuals, *Koj*: the number of applications of the item j/number of individuals, *$\overline{K}$oj*: Test length/the size of item pool, I(Θ): item information function

Table 2 shows that when examining the effect of item selection methods on the TIF, the TIF is calculated based on the standard error calculated for each individual because there is an inverse relationship between the TIF and the standard error of measurement (van der Linden, 1998). As the indicator of test efficiency value, ATL, AERI, X^2, UIR (UIR < 0.02), and OIR (OIR > 0.20) values were calculated.

Results

The results of the comparison of seven different item selection methods' TIF and test efficiency values are given below. The TIF values obtained according to item selection methods are given in Figure 3.

Figure 3

The TIF Values Obtained According to Item Selection Methods

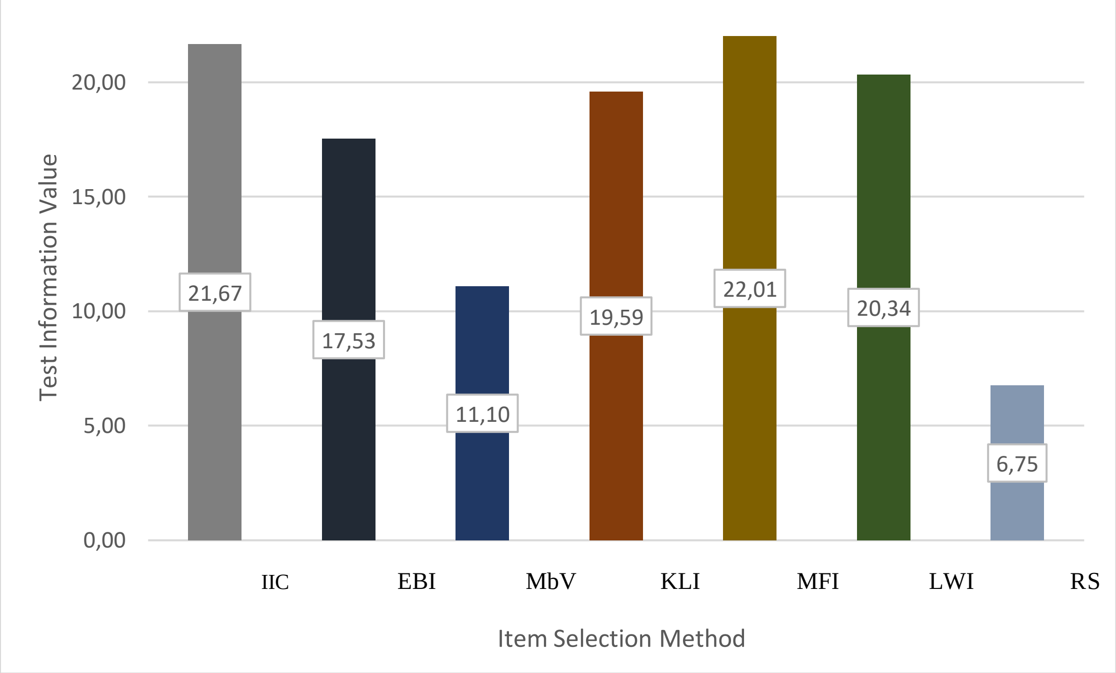


Figure 3 shows that the method which gives the highest TIF compared to other item selection methods is the MFI method (22.01) (%95 CI [22.017-21.993]. The IIC method (21.67) (%95 CI [21.661-21.682]) follows the MFI method. The method which produces the lowest test information value among the item selection methods is the RS method (6.75) (%95 CI [6.749-6.754]) which is the expected. The method which produces the lowest test information value after RS is the MbV method (11.10) (%95 CI [11.085-11.104]). The other methods which follow the MbV are EBI (17.53) (%95 CI [17.509-17.543]), KLI (19.59) (%95 CI [19.574-19.606]) and LWI (20.34) (%95 CI [20.324-20.351]) respectively. Veerkamp and Berger (1997) stated in their study that they compared the performance of the IIC, LWI and MFI methods and that the performances of these methods are close to each other as in this study.

Since the items are selected randomly in the RS method, which therefore has no adaptive test features (Han, 2012), the RS method produces the lowest test information value. The key to successful CAT implementation is choosing the best item selection method and item exposure method that will allow obtaining the best TIF (Han, 2009). It can be said that the MFI approach is used frequently in research because it is a simple, straightforward, and effective method (Han, 2009), and it was also one of the efficient methods for obtaining high test information values in this study.

The ATL, AERI, X^2, UIR and OIR values, which are indicators of test efficiency obtained according to item selection methods, are given in Table 3.

Table 3
Values for the Test Efficiency Obtained According to Different Item Selection Methods

Item Selection Method	ATL	AERI	X^2	UIR	OIR
IIC	15.00	0.03	14.94	0.05	0.81
EBI	15.00	0.03	14.94	0.03	0.73
MbV	15.00	0.03	14.94	0.00	0.51
KLI	15.00	0.03	14.94	0.06	0.83
MFI	15.00	0.03	14.94	0.05	0.79
LWI	15.00	0.03	14.94	0.05	0.80
RS	28.00	0.06	27.89	0.00	0.00

According to test efficiency values in Table 3, the ATL value is 15.00 for all item selection methods except for RS. The highest ATL value belongs to the RS method. The RS method does not produce an efficient test in terms of average test length, while other methods perform similarly. AERI represents the ratio of the total exposure rate of an item to the number of individuals. Examining the AERI values, it can be as with the ATL value, all methods perform similarly except for the RS method. As the average number of item applications was relatively higher than the others, RS (0.06) has the lowest performance.

The X^2 value represents the difference between observed and expected item use frequency. The higher this number is, the lower the test efficiency performance of the method. In other words, if this value is high, it means that some items are used repeatedly (Cheng et al., 2015). Methods other than the RS method performed close to each other, while the RS (27.89) method displayed the lowest performance as the X^2 value was higher compared to other methods. In other words, items in the item pool are used more repeatedly in the RS method compared to other item selection methods. Tested by ATL, AERI and X^2 values, it was seen that all methods except for the RS method displayed similar results. This might be related to the use of the fade-away method as the item use frequency method within the scope of this study. As Boztunç Öztürk (2014) has also stated, item use frequency ratios are close to each other since various items are used in the implementation of the fade-away method. This is supported by X^2 analysis results. Because in terms of test security, in the case the fade-away method is used as the item use frequency method, results are closer to the ideal compared to the situations where item use frequency is not controlled for (Boztunç Öztürk, 2014).

The UIR represents what percentage of items in an item pool were seen lower than 0.02 per cent. In this context, it is observed that no item (%0.00) is used in low frequency in the RS and MbV methods. Meanwhile, it is seen that the highest low exposure ratio is observed in the KLI method. The OIR represents what percentage of items in an item pool were seen higher than 0.20 per cent. In this context, the method in which the item exposure rate is highest is the KLI method, with 0.83 per cent. Following KLI is the IIC method (0.81 per cent). In the RS method (0.00 per cent), no item had a high exposure rate. In this context, it can be said that the KLI method performs in an unbalanced way in terms of underexposed and overexposed item ratios.

When low and high exposure rates are examined together, it was seen that the RS method used neither low nor high exposure items. Although the item re-use rate increases as the average test length increases in the RSI method, this increase is balanced. Among other methods with similar ATL, AERI and X^2 values, although it was seen that the KLI and IIC methods had higher low and high item exposure rates, this percentage is low. This finding indicates that those items in the pool that were used too high or too low exposure are used more frequently compared to other methods, and the item pool usage is more unbalanced, albeit very slightly (Cheng et al., 2015). This finding is also supported by the study conducted by Sulak (2013). In the relevant study, it was stated that the lowest and highest number of item usage was obtained from the KLI method, with EAP as the ability estimation method. It is thought that a similar finding was obtained in this study as the ability estimation was performed using the EAP method.

Discussion and Conclusion

The effect of various item selection methods on TIF and test efficiency in the CAT application was examined in this study. TIF is an indicator of the amount of information provided by the test and is inversely proportional to the standard error of estimation. In other words, the information provided by the test decreases as the standard error of estimation increases (van der Linden, 1998). In this context, RS is the method that gives the highest standard error, while the MFI is the method with the lowest standard error. In the study where Han (2010) compared the performances of a-stratification, IIC, LWI, KLI and GMIR methods, it was stated that the MFI, KLI and GMIR methods displayed lower standard error values compared to other methods. It was stated that these methods give higher test information, particularly in cases where the test length is shorter. Aside from this, the MFI method being an efficient method in terms of increasing the TIF is supported by the study conducted by Han (2009). It is stated that theoretically, MFI-based methods provide maximum test information with low standard error value due to choosing items that maximize the information function (Han, 2009; 2018).

In terms of ATL, AERI and X^2 values that are the indicators of test efficiency, all methods except for the RS method performed similarly. The RS method displayed a lower performance compared to other methods in terms of test efficiency. The reason for this finding might be due to the RS method not having an adaptive test feature, as items in this method are chosen randomly without any criteria (Han, 2012). In Choi and Swartz's (2009) study comparing six different item selection methods, the RS method had the lowest correlation between the observed and expected ability values. Another situation related to the use of an item pool is that when the KLI, MFI and LWI methods which show similar performances are examined, items with higher discrimination were selected more in implementation. A similar finding can be found in a study conducted by Sulak (2013). Although it is stated that one of the ways to prevent the use of items with high discrimination is the a-stratification method (Sulak, 2013, Yi & Chang, 2003), this study did not take into account the a-stratification method. A general evaluation of the performance of methods shows that the methods other than RS have no advantage over one another both in terms of the ATL, AERI and X^2 values.

Evaluating the low and high exposure percentages of the items, it is seen that these values are lower than one and close to each other. In the study where Han (2009) compared different item selection methods, it was stated that in cases where the fade-away item use frequency is employed, the item exposure rate was more under control. The reason for low and high item exposure percentages being very low and close to each other in this study might be due to the use of the fade-away item use frequency method. In terms of item pool use frequency, MFI, a- stratification, the fact that LWI and KLI methods perform similarly is supported by the work of Sulak (2013). It was seen that the item's high exposure percentage was high in the KLI and IIC methods. Although this percentage is not very high, it poses a threat to test security because it means some items are shown to individuals at a far higher rate due to the unbalanced use of the item pool. This is a factor that threatens test security. For this reason, although it is a low ratio under the conditions taken into account in this study, the use of KLI and IIC methods should be approached with caution due to concerns about item pool security.

The purpose of the CAT applications is to measure the subject with a high validity-reliability and high-test information at the ability level of each individual. In this context, CAT is a tool to produce high test information and efficient tests suitable for the ability level of each individual. It may be suggested that, in terms of the item selection method, the MFI method performs better compared to other methods, and the RS method has the worst performance and should not be selected because the RS method also has not an adaptive feature. The MbV method followed the RS method in the aspect of worse performance in TIF. Since methods other than the RS and MbV display similar performances in terms of TIF and test efficiency, any one of them can be chosen. However, in order to obtain better results in terms of TIF, it is suggested that implementers use the MFI method. For further studies, it may be suggested that as this study worked with a constant sample and item pool size, they should examine the effects of these methods in a smaller or larger sample and item pool sizes. Aside from these, a constant start rule, termination rule, ability estimation method and item use frequency was employed in the CAT implementation process in this study. The effects of item selection methods on ability estimations and test efficiency can be examined by modifying the specified CAT conditions.

Declarations

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

Ethical Approval: The data used in this study were generated by simulation. Therefore, ethical approval is not required.

References

Babcock, B. & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36*(7), 565- 580. https://doi.org/10.1177/0146621612455090

Babcock, B. & Weiss, D.J. (2012). Termination criteria in computerized adaptive tests: Do variable-length CAT's provide efficient and effective measurement? *International Association for Computerized Adaptive Testing, 1*, 1-18. http://dx.doi.org/10.7333%2Fjcat.v1i1.16

Baker, F. (1986). The basics of item response theory. *Journal of Educational Measurement, 23*(3), 267-270.

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. Marcel Dekker.

Balta, E., & Uçar, A. (2022). Investigation of measurement precision and test length in computerized adaptive testing under different conditions, *E-International Journal of Educational Research, 13*(1), 51-68. https://doi.org/10.19160/e-ijer.1023098

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chaps. 17–20). AddisonWesley.

Blais, J. & Raiche, G. (2010). Features of the sampling distribution of the ability estimate in Computerized Adaptive Testing according to two stopping rules. *Journal of applied measurement, 11*(4), 424-31.

Bock, R. D. & Aitkin, M. (1981).Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*(4), 443–459. https://link.springer.com/article/10.1007/BF02293801

Bock, R. D. & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, 6(4), 431– 444. https://doi.org/10.1177/014662168200600405

Boyd, M. A. (2003). *Strategies for controlling testlet exposure rates in computerized adaptive testing systems* [Unpublished Doctoral Thesis]. The University of Texas.

Boztunç Öztürk, N. (2014). *Bireyselleştirilmiş bilgisayarlı test uygulamalarinda madde kullanım sıklığı kontrol yöntemlerinin incelenmesi [Investigatıon of item exposure control methods in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Boztunç Öztürk, N. & Doğan, N. (2015). Investigating item exposure control methods in computerized adaptive testing. *Educational Sciences: Theory and Practice, 15*(1), 85-98. https://doi.org/10.12738/estp.2015.1.2593

Brown, A. (2018). Item response theory approaches to test scoring and evaluating the score accuracy. In Irwing, P., Booth, T. & Hughes, D. (Eds.), *The Wiley Handbook of Psychometric Testing*. John Wiley & Sons.

Chang, H.-H. & Ying, Z. (1999). a-stratified multistage computerized adaptive testing. *Applied Psychological Measurement, 23*(3), 211-222. https://doi.org/10.1177/01466219922031338

Choi, S. W. & Swartz, R. J. (2009). Comparison of CAT item selection criteria for polytomous items. *Applied Psychological Measurement, 33*(6), 419–440. https://doi.org/10.1177/0146621608327801

Cheng, Y., Patton, J.M. & Shao, C. (2015). a-Stratified computerized adaptive testing in the presence of calibration error. *Educational and Psychological Measurement, 75*(2), 260-283. https://doi.org/10.1177/0013164414530719

Costa, D., Karino, C., Moura, F. & Andrade, D. (2009, June). *A comparision of three methods of item selection for computerized adaptive testing* [Paper Presentation] The meeting of 2009 GMAC Conference on Computerized Adaptive Testing. Retrieved from www.psych.umn.edu/psylabs/CATCentral/

Çıkrıkçı-Demirtaşlı, N. (1999). Psikometride yeni ufuklar: Bilgisayar ortamında bireye uyarlanmış test [New horizons in psychometrics: Individualized test in computer environment]. *Türk Psikoloji Bülteni, 5*(13), 31-36.

Davis, L. L. (2002). *Strategies for controlling item exposure in computerized adaptive testing with polytomously scored items.* [Unpublished Doctoral Dissertation], The University of Texas.

Davis, L. L., & Dodd, B. G. (2005). *Strategies for controlling item exposure in computerized adaptive testing with partial credit model.* Pearson Educational Measurement Research Report 05-01.

Deng, H., Ansley, T. & Chang, H. (2010). Stratified and maximum information item selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*(2), 202-226. https://www.jstor.org/stable/20778948

Doğan, C.D. & Aybek, E.C. (2021*). R-Shiny ile psikometri ve istatistik uygulamaları [Psychometric and statistical applications with R-Shiny]*. Pegem Akademi.

Eggen, T.J.H.M. (2001). Overexposure and underexposure of items in computerized adaptive testing. Measurement and Research Department Reports, 2001-1. Citogroep

Eggen, T.H.J.M. (2004). *Contributions to the theory and practice of computerized adaptive testing*. Print Partners Ipskamp B.V., Citogroup Arnhem.

Eggen, T.H.J.M. (2012). Computerized adaptive testing item selection in computerized adaptive learning systems. *Psychometrics in Practice at RCEC*, 11.

Eroğlu, M.G. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında farklı sonlandırma kurallarının ölçme kesinliği ve test uzunluğu açısından karşılaştırılması [Comparison of different test termination rules in terms of measurement precision and test length in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement*, *6*(1), 109–127.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational measurement*, *21*(4), 347-360. https://www.jstor.org/stable/1434586

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* SAGE Publications.

Han, K.T. (2009). *Gradual maximum information ratio approach to item selection in computerized adaptive testing*. Council Research Reports, Graduate Management Admission.

Han, K.T. (2010). *Comparison of Non-Fisher Information Item Selection Criteria in Fixed Length Computerized Adaptive Testing* [Paper Presentation] The Annual Meeting of the National Council on Measurement in Education, Denver.

Han, K. T. (2011). *User's Manual: SimulCAT.* Graduate Management Admission Council.

Han, K.T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement*, *36*(1), 64-66.

Han, K. (2018). Components of item selection algorithm in computerized adaptive testing. *J Educ Eval Health Prof*, *15*(7). https://doi.org/10.3352/jeehp.2018.15.7

Kaptan, F. (1993). *Yetenek kestiriminde adaptive (bireyselleştirilmiş) test uygulaması ile geleneksel kağıt-kalem testi uygulamasının karşılaştırılması [Comparison of adaptive (individualized) test application and traditional paper-pencil test application in ability estimation]* [Unpublished Doctoral Dissertation]. Hacettepe University

Keller, A.L. (2000). *Ability estimation procedures in computerized adaptive testing*. Technical Report, American Institute of Certified Public Accountants-AICPA Research Concortium-Examination Teams.

Kezer, F. (2013). *Bilgisayar ortamında bireye uyarlanmış test stratejilerinin karşılaştırılması [Comparison of computerized adaptive testing strategies]* [Unpublished Doctoral Dissertation]. Ankara University.

Kingsbury, G. G. & Zara, A. R. (1989). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6

Lee, H., & Dodd, B. G. (2012). Comparison of exposure controls, item pool characteristics, and population distributions for CAT using the partial credit model. *Educational and Psychological Measurement, 72*(1), 159-175. https://doi.org/10.1177/0013164411411296

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

Moyer, E. L., Galindo, J. L., & Dodd, B. G. (2012). Balancing flexible constraints and measurement precision in computerized adaptive testing. *Educational and Psychological Measurement, 72*(4). https://doi.org/10.1177/0013164411431838

Ranganathan, K. & Foster, I. (2003). Simulation studies of computation and data scheduling algorithms for data grids. *Journal of Grid Computing, 1*, 53-62. https://doi.org/10.1023/A:1024035627870

Risk, N.M. (2010). *The impact of item parameter drift in computer adaptive testing* (CAT) [Unpublished doctoral dissertation]. University of Illinois.

Rudner, L.M. & Guo, F. (2011). Computer adaptive testing for small scale programs and instructional systems. *Graduate Management Council (GMAC),* 11(01), 6-10.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika monograph supplement*, *34*(4). https://doi.org/10.1002/j.23338504.1968.tb00153.x

Sulak, S. (2013). *Bireyselleştirilmiş bilgisayarlı test uygulamalarında kullanılan madde seçme yöntemlerinin karşılaştırılması [Comparision of item selection methods in computerized adaptive testing]* [Unpublished Doctoral Dissertation]. Hacettepe University.

Sulak, S. & Kelecioğlu, H. (2019). Investigation of item selection methods according to test termination rules in CAT applications. *Journal of Measurement and Evaluation in Education and Psychology*, *10*(3), 315-326. https://doi.org/10.21031/epod.530528

Stahl, J. A. & Muckle, T. (2007, April). *Investigating displacement in the Winsteps Rasch calibration application* [Paper Presentation] The Annual Meeting of the American Educational Research Association, Chicago, IL.

Stocking, M. L. (1992). *Controlling item exposure rates in a realistic adaptive testing paradigm*. Research Report 93-2, Educational Testing Service.

Şahin, A. (2012). *Madde tepki kuramında test uzunluğu ve örneklem büyüklüğünün model veri uyumu, madde parametreleri ve standart hata değerlerine etkisinin incelenmesi [An investigation on the effects of test length and sample size in item response theory on model-data fit, item parameters and standard error values]* [Unpublished Doctoral Dissertation]. Hacettepe University.

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

Urry, V. (1977). Tailored testing: A successful application of latent trait theory. *Journal of Educational Measurement, 14*, 181-196. https://www.jstor.org/stable/1434014

van der Linden, W. (1998). Bayesian item selection criteria for adaptive testing. *Psychometrika*, 63(2), 201–216. https://doi.org/10.1007/BF02294775

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Some New Item Selection Criteria for Adaptive Testing. *Journal of Educational and Behavioral Statistics*, *22*(2), 203-226. https://doi.org/10.3102/10769986022002203

Wainer, H. (1993). Some practical considerations when converting a linearly administered test to an adaptive format. *Educational Measurement: Issues and Practice*, *12*(1), 15–20. http://dx.doi.org/10.1111/j.1745-3992.1993.tb00519.x

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450. https://doi.org/10.1007/BF02294627

Weiss, D. J. (1982). Latent Trait Theory and Adaptive Testing. In David J. Weiss (Ed.). *New horizons in testing: Latent trait test theory and computerized adaptive testing* (pp. 5-7). Academic Press.

Weiss, D. J. & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361–-375. https://www.jstor.org/stable/1434587

Wen, H., Chang, H. & Hau, K. (2000). *Adaption of a-stratified Method in Variable Length Computerized Adaptive Testing*. American Educational Research Association Annual Meeting, Seattle.

Yi, Q. & Chang, H. (2003). a-Stratified CAT design with content blocking. *British Journal of Mathematical and Statistical Psychology*, *56*, 359–378. https://doi.org/10.1348/000711003770480084

Zwinderman, A. H., & van den Wollenberg, A. L. (1990). Robustness of marginal maximum likelihood estimation in the Rasch model. *Applied Psychological Measurement*, *14*(1), 73-81. https://doi.org/10.1177/014662169001400107