HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Educational Sciences
Educational Measurement and Evaluation Program

THE EFFECT OF ITEM SELECTION AND PARAMETER ESTIMATION METHODS TO THE ACCURACY OF PRETEST ITEM PARAMETERS ON ONLINE CALIBRATION IN CAT

Levent ERTUNA

Ph.D. Dissertation

Ankara, 2020

With leadership, research, innovation, high quality education and change,

To the leading edge... Toward being the best...

HACETTEPE ÜNİVERSİTESİ
EĞİTİM BİLİMLERİ ENSTİTÜSÜ

Department of Educational Sciences

Educational Measurement and Evaluation Program

THE EFFECT OF ITEM SELECTION AND PARAMETER ESTIMATION METHODS TO THE ACCURACY OF PRETEST ITEM PARAMETERS ON ONLINE CALIBRATION IN CAT

MADDE SEÇİM VE PARAMETRE KESTİRİM YÖNTEMLERİNİN BBT ONLİNE KALİBRASYONDA ÖN TEST MADDE PARAMETRE KESİNLİĞİNE ETKİSİ

Levent ERTUNA

Ph.D. Dissertation

Ankara, 2020

Abstract

Computerized adaptive test (CAT) has a possible risk that the quality of the item bank decreases over time due to the exposed items. The best possible and advantageous solution to this is to implement the online calibration. This study was aimed to investigate the effect of online calibration components on precision in parameter estimation and the cumulative sample size (specified to the calibration method). It was also aimed to transfer Joint Maximum Likelihood as a pretest calibration method to the online calibration procedure and assess this method's feasibility. The simulation study was conducted under one-parameter logistic (1-PL) and two-parameter logistic (2-PL) model to compare the pretest item selection methods (Maximum Fisher Information-MFI, D-optimal value design-DVOD, and Bayesian D-optimal design-BDOD), the parameter estimation methods (Joint Maximum Likelihood-JML and Marginal Maximum Likelihood with One EM Cycle-OEM), the sample size of the random calibration stage (250, 500, and 1000) and the calibration sample size of per pretest item (250, 500, and 1000). The performance of these factors on the parameter precision was evaluated by calculating bias and root mean squared error (RMSE). The results indicate that the performances of item selection methods differ according to Item Response Theory (IRT) models and the parameter estimation methods. Among the calibration methods, OEM has successfully estimated the most precise item parameters although JML performed better in some conditions. The sample size of the random stage did not have a characteristic effect on parameter estimation. Lastly, the parameter accuracy gets higher as the calibration sample size increases.

Keywords: item response theory, computerized adaptive testing, online calibration, pretest item, field-test item, item parameter estimation, pretest item selection method

Öz

Bireyselleştirilmiş Bilgisayarlı Testler (BBT)'de maddelerin teşhir olmasından kaynaklı olarak madde havuzunun kalitesinin düşme riski vardır. Bunun için en iyi ve en avantajlı çözüm, online kalibrasyon prosedürünün uygulanmasıdır. Bu çalışmanın amacı online kalibrasyon bileşenlerinin parametre kesinliğine ve madde seçim yöntemi özelinde kümülatif örneklem büyüklüğüne etkisini incelemektir. Ayrıca çalışmanın diğer bir amacı da Ortak Maksimum Olabilirlik yönteminin parametre kestirim yöntemi olarak online kalibrasyon prosedürüne uygulamak ve uygunluğunun değerlendirilmesidir. Bir Parametreli Lojistik (1-PL) model ve İki Parametreli Lojistik (2-PL) model altında öntest madde seçim yöntemlerini (Maksimum Fisher Yöntemi, D-optimal Değer Deseni, Bayesyen D-optimal Deseni), parametre kestirim yöntemlerini (Ortak Maksimum Olabilirlik ve Marjinal Maksimum Olabilirlik ile Tek Beklenti Maksimizasyon Döngüsü), seçkisiz kalibrasyon aşaması örneklem büyüklüklerini (250, 500 ve 1000) ve her öntest maddesi için örneklem büyüklüklerini (250, 500 ve 1000) karşılaştırmak amacıyla simülasyon çalışması yapılmıştır. Bu faktörlerin parametre doğruluğu üzerindeki performansları yanlılık ve Hata Kareleri Ortalamalasının Karekökü (HKOK) hesaplanarak değerlendirilmiştir. Sonuçlar madde seçim yöntemlerinin performanslarının Madde Tepki Kuramı (MTK) modellerine ve parametre kestirim yöntemlerine göre farklılaştığını göstermektedir. Kalibrasyon yöntemleri arasında Ortak Maksimum Olabilirlik yöntemi bazı koşullarda daha iyi performans göstermesine rağmen Tek Beklenti Maksimizasyon Döngüsü madde parametrelerini en doğru şekilde başarıyla kestirmiştir. Seçkisiz aşama örneklem büyüklüğünün parametre kesinliği üzerinde karakteristik bir etkiye sahip olmaması elde edilen diğer bir sonuçtur. Son olarak parametre kesinliği kalibrasyon örneklem büyüklüğünün artmasıyla yükselmiştir.

Anahtar sözcükler: madde tepki kuramı, bireyselleştirilmiş bilgisayarlı test, online kalibrasyon, öntest maddesi, parametre kestirim yöntemi, öntest madde seçim yöntemi

Acknowledgements

First of all, I would like to express my sincere gratitude to my advisor, Assoc. Prof. Dr. Burcu Atar for her guidance, support, advice, encouragement, and more importantly, her kindness and patience throughout the research. I am grateful for her fast feedback on my drafts and proofreading.

I would extend my thanks to my committee member, Prof. Dr. Neşe Güler for her comments and suggestions. She has always supported me with all her sincerity. I would also like to thank my other committee members, Prof. Dr. Hakan Yavuz Atar, Prof. Dr. Nuri Doğan, and Assist. Prof. Dr. Derya Çobanoğlu Aktan for providing their valuable contributions to improve my dissertation.

I would like to pay my special regards to the members of the division of Educational Measurement and Evaluation, Prof. Dr. Selahattin Gelbal and Prof. Dr. Hülya Kelecioğlu for their contributions to my development during my Ph.D. education.

I would like to thank the Scientific and Technological Research Council of Turkey (TÜBİTAK) for their scholarships under the programmes of 2211-A (National PhD Scholarship) and 2214-A (International Research Fellowship).

My sincere gratitude extends to Prof. Dr. Cees Glas for working me together on my project at Research Methodology, Measurement and Data Analysis Department (OMD) of University of Twente in the Netherlands. His enthusiasm, motivation, knowledge, and perspective have always inspired me to work. I am always grateful to him for his practical solutions, support, and advice.

I would like to appreciate my colleagues; Assist. Prof. Dr. Gülden Kaya Uyanık and Assist. Prof. Dr. Süleyman Demir, and my friends.

Last but not least, my deepest thanks go to my family and my wife's family for their continuous support. I would like to express my heartfelt gratitude to my bellowed wife, Tülay for her unconditional supports, constant love, belief in me, and encouragement.

Table of Contents

List of Tables

List of Figures

Symbols and Abbreviations

BDOD: Bayesian D-Optimal Design

CAT: Computerized Adaptive Test

DVOD: D-Optimal Value Design

EM: Expectation Maximization

IRT: Item Response Theory

JML: Joint Maximum Likelihood

MCAT: Multidimensional CAT

MEM: Marginal Maximum Likelihood Estimation with multi EM cycle

MFI: Maximum Fisher Information

N: The Calibration Sample Size of Per Pretest Item

n_r: The Sample Size of Random Calibration Phase

OEM: Marginal Maximum Likelihood Estimation with one EM cycle

RMSE: Root Mean Squared Error

UCAT: unidimensional CAT

WLE: Weighted Maximum Likelihood Estimator

1-PL: The One-Parameter Logistic (model)

2-PL: The Two-Parameter Logistic (model)

3-PL: The Three-Parameter Logistic (model)

Chapter 1
Introduction

The tests used for many different purposes in the range from low-stakes to high-stakes must fulfil some basic standards associated with reliability, validity, and fairness of their scores. In order to meet these standards, the psychometric developments (such as Item Response Theory-IRT) and the advantages of technology had been utilized and this brought along the change from the linear test fashion to the adaptive test fashion. Computerized adaptive test (CAT) is one of the best practices of adaptive test which aims to select and sequentially administer the most appropriate items to the examinees according to their provisional abilities. In contrast to traditional tests, CAT estimates the abilities more precisely with fewer items. Thanks to these features, it has expanded its popularity and has been used for wide range of purposes (i.e., large-scale assessment, college placement, achievement testing, monitoring education, and higher-education admissions) with various areas (i.e., education, psychology, health-outcome, and even surprisingly marketing) (van der Linden & Glas, 2010). In addition, it has been an increasingly important research area in psychometry and educational testing since 1970s (Luecht & Sireci, 2011).

With the tailored nature of CAT, it provides many advantages to test takers and test designers. It estimates the ability with less measurement error by presenting more suitable items close to the examinee's ability level. This means that test scores are more reliable. In this way, it enables the examinee to maintain their test motivation and prevents contamination of unrelated variables such as test anxiety, examinee fatigue to test scores. It can also balance the content with the help of the item selection algorithm for the purpose of the test. Another advantage of CAT is that it allows great flexibility for testing time and location. As a result of assigning different sets of items to each test taker, it increases test security. It also allows the presentation of multimedia-based items that cannot be used in traditional paper and pencil tests. It gives test takers immediate access to feedback, reports, and scores of their test (Linacre, 2000; Mead & Drasgow, 1993; Rudner, 1998; Wainer & Eignor, 2000; Weiss, 1982). Despite CAT provides numerous advantages as mentioned above, it has some disadvantages and limitations. It does not allow the examinee to skip an item, to go back to the previous item, and to review the former responses. CAT is also a costly process. Although it provides tests of shorter lengths, it needs to have an item bank wide enough to span all ability levels in order to ensure its adaptive nature. Item exposure is more of a concern in CAT administration (Colvin, 2014). Therefore, in the next step, it raises the problem of maintaining the item bank. In addition, it is costly to access to computer networks and to have computer hardware and software that can enable the continuity of CAT (Luecht & Sireci, 2011; Mills & Stocking, 1995; Vispoel, Rocklin, & Wang, 1994; Wainer, 1993).

CAT system is performed with five integrated components; a calibrated item pool/bank consisting of items with predetermined parameters, starting rule that initiates the procedure by the selecting the first item, item selection algorithm that determines the suitable items, the scoring method that estimates both interim and final ability of examinee and the stopping rule that determines how to terminate the test (Weiss & Kingsbury, 1984; Thomson, 2007; Thompson & Weiss, 2011). In order to realize its advantages, CAT should be administered to each examinee with different ability levels with higher quality items. That is, the item bank should have sufficient number of items targeted at each ability level according to the purpose of the test. Therefore, the item bank quality has been thought of as a key factor as it is directly related to the success of CAT (e.g. Flaugher, 2000; Wise & Kingsbury; 2000; Thompson & Weiss, 2011).

Statement of the Problem

The item bank is developed within a general plan starting with the steps of writing and reviewing items. After that, in order to estimate the parameters of these items, the pretesting session is carried out by administering the items to a large number of the examinee with the paper and pencil form. After the rules related to the components of CAT are decided, the CAT test is published. Due to the fact that CAT is a continuous system applied to different examinees at different time intervals, the critical issue that should be considered as the last step is the continuity of CAT (Flaugher, 2000; Thompson & Weiss, 2011). This is directly related to maintaining the item bank since some items in the bank are overexposed or outdated over time. In this case, these items should be retired from the bank and replaced by new ones (Wainer & Mislevy, 1990). The entire process is called "item replenishment” or “refreshing item bank”. To do this, these additional items must be calibrated as in the first time the items bank is structured and be transferred to item bank for operational use. At this point, the efficient operation of the calibration process is critical. The precision of the item parameter estimation is known to directly affect the validity and reliability of the test scores, and is essential for test scores, test equating, differential item functioning. Therefore, it is crucial that item parameters can be estimated on large scale, efficiently and economically (Stout, Ackerman, Bolt, Froelich, & Heck, 2003).

There are two methods to handle the parameter estimation of new items (or field-test item) for item replenishment in CAT. The first calibration method, conventional offline calibration, is similar to the method used for the initial development of item bank, is based on the anchor item design. In this method, new items are assigned to a group of volunteers with using a paper and pencil form, and then are calibrated using these obtained data. Finally, the linking process is employed to ensure that the parameters of the new items are on the same scale as the parameters of the operational items (Wainer & Mislevy, 1990). In this method, it may not always be feasible to perform the calibration with large samples. One of the potential disadvantages of this method is that the data obtained are unreliable due to the volunteers’ low motivation during test. In this respect, this method is both expensive and time-consuming (Chen & Wang, 2016; He, Chen, & Li, 2020). In the second method, named as online calibration, new items are carefully embedded among operational items and are “seeded” inconspicuously to the examinees during their active testing in the CAT scenario (Ban, Hanson, Wang, Yi, & Harris, 2001; Stocking, 1988; Wainer & Mislevy, 1990).

The motivation for the emergence of this method is very similar to the idea of operating the CAT system meaning that the examinees’ abilities are estimated more effectively by administrating them appropriate set of items adaptively. Similarly, online calibration is based on the idea that the parameters of new items (called pretest items in online calibration context) can be precisely estimated by giving an optimal batch of the examinees (Ali & Chang, 2014; Zheng, 2014). In this context, online calibration is operated in as follows. During the examinee’s operational test, a certain number of items are administrated to the examinee at predetermined position in the test (seeding position/location) by selecting from the pretest bank which is a collection of pretest items. Firstly, since the parameters of the pretest items are not estimated at all, the responses of the active candidates to these items are recorded and stored. At this stage, the pretest items are randomly selected from the pretest item bank. When these pretest items reach a certain number of responses (as defined the sample size of the random calibration stage) with the continuation of operational CAT, the initial parameters of these items are estimated. In the following stages, pretest items are administrated to the examinee according to the determined rules at the end of the operational CAT test. With the pretest item response of the active examinees, the parameters of that item are updated by re-estimated using the responses of the previous examinees recorded to the same item. The procedure is continued until the specified stopping criterion is met (Zheng, 2014). At the end of this process, all pretest items have final parameters.

Online calibration has many clear advantages over the traditional method. As the pretest items are presented unnoticed to the examinees, they can maintain their test mode and can respond to the items with this motivation (Parshall, 1998). This ensures that the obtained data in this method is more reliable than the traditional method. This procedure makes it possible to estimate the test takers' abilities and the pretest item parameters simultaneously. In online calibration procedure, pretest item parameters are located on the same scale as operational item parameters, thus further complex techniques such as equating and linking are not needed. With all these advantages, online calibration is cost-effective and time-saving (Chen, Xin, Wang, & Chang, 2012; Makransky & Glas, 2010).

The online calibration procedure consists of two fundamental elements; online calibration design (as denoted pretest item selection method) and online calibration methods (as denoted item parameter estimation method). Online calibration design is related to the selection rules of the sample for the pretest items and more generally it refers to the pretest item selection rules or methods. The calibration method describes the parameter estimation methods used in this process specific to online calibration. In addition to those fundamental elements, the online calibration procedure consists of three administration elements; seeding position, termination rule and sample size as a type of termination rule. The seeding position, as the name suggests, describes where the pretest items will be administrated during the active test. The stopping rule describes how the online calibration procedure will be finished. The sample size is defined the number of examinees required to complete parameter estimation for a pretest item (as defined the calibration sample size of per pretest item) or all pretest items (as defined the cumulative sample size). Previous studies have examined these elements on parameter estimation in different CAT designs; unidimensional CAT (UCAT) (e.g. Ali & Chang, 2014; Ban et al., 2001; Chang & Lu, 2010; He, 2015; He, Chen, Li & Zhang, 2017; He et al., 2020; Kingsbury, 2009; Lu, 2014; Makransky & Glas, 2010; van der Linden & Ren, 2015; Zheng, 2014; Zheng, 2016; Zheng & Chang, 2017) and multidimensional CAT (MCAT) (e.g. Chen & Wang, 2016; Chen, 2017; Chen et al., 2017; Zheng, 2014; Zheng, 2016; Zheng & Chang, 2017). However, few of these studies (e.g. He et al., 2017; He et al., 2020; Zheng, 2014; Zheng & Chang, 2017) have attempted to the examination of these components together in UCAT. Without conducting similar studies to these studies, it will not be known how operating the online calibration procedure will affect the accuracy of pretest item parameter and therefore it is clear that the item parameters will be estimated with more measurement errors.

Furthermore, the obtained data in CAT is in a sparse and based restricted range of examinees' ability due to its structure. This causes a change in the parameter estimation of the pretest item. Therefore, parameter estimation is more complex than traditional methods (Ban et al., 2001; Hsu, Thompson, & Chen, 1998; Chen, 2017). In order to overcome this problem, a number of parameter estimation methods have been suggested: Method A (Stocking,1988), Method B (Stocking,1988), Marginal Maximum Likelihood Estimation with one expectation maximization (EM) cycle (OEM) method (Wainer & Mislevy, 1990), Marginal Maximum Likelihood Estimation with multiple expectation maximization (EM) cycle (MEM) method (Ban et al., 2001), BILOG/Strong Prior method (Ban et al., 2001), and maximum likelihood estimation-Lord's bias-correction with iteration – Method A method (He et al., 2017). These methods are generally proposed by adapting traditional item response theory (IRT) parameter estimation methods for online calibration. Due to the importance of parameter estimation, this element is a complex but fruitful area for the adaptation and transfer of traditional methods and development of new methods.

As mentioned above, although there are several studies on online calibration, this field still needs to be improved. This study was designed based on all these problem situations and potential positive contributions.

Aim and Significance of the Study

The current study had two key aims. Firstly, it was aimed to investigate the effect of the pretest item selection methods, the parameter estimation methods, the sample size of the random calibration stage and the calibration sample size of per pretest item on precision in parameter estimation. It was also aimed to assess the effect of the pretest item selection methods on the cumulative sample size. Secondly, it was aimed to use Joint Maximum Likelihood as a pretest item parameter estimation method in the online calibration procedure and assess its feasibility. In order to evaluate the effectiveness of parameter estimation, bias and root mean squared error (RMSE) were calculated.

In the CAT system, it is a possible risk that the quality of the item bank decreases over time due to the exposed or obsolete items. Precautions should be taken to prevent potential misconduct and decisions that may cause this risk to the continuity of CAT. Therefore, item replenishment that refers to the replacement of these items with new ones is required to maintain the item bank. The best possible and advantageous solution to this is to carry out the online calibration procedure. The key issue in this is the precise estimation of item parameters, that is to say, it is the estimation of the new items’ parameters with the least errors and the optimum number of samples. For this purpose, it is necessary to know the effects of online calibration elements on the item parameter recovery and thus the procedure can be run at the optimum sample size using the best proper item selection design, the best performing parameter estimation method. This aspect of the study serves and contributes to this purpose.

The literature about online calibration begins with Stocking’s (1988) study. The previous studies of this issue (e.g. Ali & Chang, 2014; Ban et al., 2001; Chang & Lu, 2010; He et al., 2017; He et al., 2020; Kingsbury, 2009; Makransky & Glas, 2010; van der Linden & Ren, 2015; Zheng, 2014; Zheng, 2016; Zheng & Chang, 2017) have generally focused on online calibration design and related item selection methods, parameter estimation methods, sample size, and their combination. There have been a small number of published studies (e.g. Stocking, 1988; Wainer & Mislevy, 1990; Ban et al., 2001; He et al., 2017) among these studies proposed the pretest item parameter estimation method for unidimensional CAT. Up to now, except for one study (Verschoor, Berger, Moser, & Kleintjes, 2019), no research has been found that examined the performance of JML as a parameter estimation method. This study provides information about the feasibility of this simple method in the online calibration procedure and shows its possible advantages and disadvantages that may arise within the design of this study. It also enabled the fundamental and administration elements of online calibration (the sample size of the random calibration stage, item selection methods, parameter estimation methods, and the calibration sample size of per pretest item) to explore together. The present study provided an opportunity to assess these conditions not only the parameter accuracy but also the cumulative sample size in online calibration. Moreover, although there have been studies on CAT in Turkey (i.e., Aybek & Demirtaslı, 2017; Bulut & Kan, 2012; Özberk & Gelbal, 2017), there is no study on online calibration. Therefore, it should make an important contribution to the national and international literature about online calibration.

Research Questions

What is the effect of the sample size of the random calibration stage, item selection methods, parameter estimation methods, and the calibration sample size of per pretest item on precision in parameter estimation and cumulative sample size in online calibration procedure?

Sub research questions. The sub research questions are as follows.

1. What is the effect of different item selection methods (Maximum Fisher Information method and D-optimal value design, and, Bayesian D-optimal design) on precision in parameter estimation and cumulative sample size in online calibration procedure?
2. What is the effect of different parameter estimation methods (Joint Maximum Likelihood and Marginal Maximum Likelihood with One Expectation Maximization Cycle) on precision in parameter estimation in online calibration procedure? Can Joint Maximum Likelihood be used as a parameter estimation method in the online calibration procedure?

3. What is the effect of the sample size of the random calibration stage (250, 500, and 1000) on precision in parameter estimation in online calibration procedure?
4. What is the effect of the calibration sample size of per pretest item (250, 500, and 1000) on precision in parameter estimation in online calibration procedure?

Limitations

For this study, three limitations need to be considered. First, the online calibration procedure is carried out separately according to 1-PL, and 2-PL IRT models. However, the study did not evaluate the 3-PL model. Second, the previous studies have been suggested several item selection designs/methods (called methods in next sections) and item calibration methods for online calibration. This study is unable to encompass the entire pretest item selection methods and item calibration methods. It only uses Maximum Fisher Information (MFI) method and D-optimal value design, and, Bayesian D-optimal design as pretest item selection methods and Joint Maximum Likelihood (JML), and Marginal Maximum Likelihood with One EM Cycle (OEM) as parameter estimation methods. Lastly, in some studies in the online calibration literature, the running time of simulation is considered as a criterion to compare the item calibration methods. Although it was recorded in this study, it was not used as a criterion because the simulation studies were carried out on different computers with different hardware.

Chapter 2
Literature Review

Computerized Adaptive Testing

Computerized adaptive testing (CAT) is a modern test administration design as an alternative to linear test design such as fixed-length conventional paper-pencil tests. CAT is used to measure the latent trait of an examinee such as ability, personality, attitude by taking advantage of technology. The integration of CAT with item response theory makes it possible to estimate the ability level of an examinee more effectively with a shorter test. In CAT, the items are sequentially administrated to an examinee and this process is performed adaptively meaning that an item is selected according to the examinee’s provisional ability level estimation based on the response to the previous item (Lord, 1980a; Weiss & Kingsbury, 1984).

As a result of both the adaptive nature of CAT and the use of computer and internet technology, it has many advantages. The first and one of the most crucial of these, as mentioned above, is that the ability can be estimated with fewer measurement errors than the paper-pencil tests since the examinees receive the appropriate items for their abilities. Accordingly, this could make it possible to have shorter tests and less time (by up to %50) without renouncing the abilities precision. Due to the continuous practice, it has testing flexibility that enables taking the test at the preferred time and location. Thanks to the adaptive item selection and hence each examinee receives different tests, the risk of test fraud and cheating is minimized and it preserves the test fairness. With the help of computer technology, it not only allows the use of different item formats (multiple media) as opposed to the traditional item formats but also provides speedy feedback to the examinee (Linacre, 2000; Rudner, 1998; van der Linden & Glas, 2010; Wainer & Eignor; Weiss, 1982).

In the literature, it is stated that the components required to operate the CAT system are as follows; an item bank with the known item parameters, the item selection methods for both first item/s and next items, the ability estimation method for both interim and final ability and the termination rule (Magis & Raîche, 2012; Reckase, 1989; Thompson & Weiss, 2011; Weiss & Kingsbury, 1984). Using these components, the CAT algorithm runs the process in four steps: initial, test, stopping and final. First, the first item is selected and given to the examinee from the item bank in the initial step. In the test step, the provisional ability is estimated using the ability estimation criteria and the item selection rule is activated. The operations in this step are repeated until the termination rule is met. In the stopping step, the test is terminated according to the pre-defined rules. In the final step, the final ability of the examinee and its measurement error is estimated by using the ability estimation criteria (Magis & Raîche, 2012; Magis, Yan, & Von Davier, 2017).

The item bank is an accumulation of operational items to be administrated to examinees in CAT. It is a central component that ensures the continuity of the CAT process. It should have a sufficient number of pre-administrated, calibrated, and ready for use items to cover the entire range of ability (Wainer & Dorans, 2000). In order to develop the item bank, the properties of the test are determined according to the purpose of the test and then the items representing each content are written, reviewed for the fairness and the quality. After this stage, the pretesting stage is carried out and the parameters of the items are estimated. The parameters and statistical properties of the pretest are evaluated according to the determined criteria (IRT assumption, model fit statistic, et cetera) if any, the items which do not have the desired properties are eliminated. To ensure the continuation of the item bank, it should be checked periodically in terms of item exposure rate, item drift, and content balance (Magis et al., 2017).

The first step of CAT involves selecting at least one item and administer it to the examinee to starting the test. This stage is operated in different ways depending on whether having any prior information about the examinee’s ability level. Mostly, this information is not available and the administration starts by clearly selecting the medium-difficult or the most informative item. In this case, the initial θ level is mostly fixed to zero (Magis et al., 2017). If this information is available, the item whose difficulty close to the θ level of the examinee is selected and given according to this information. This can also shorten the test length (Thissen, & Mislevy, 2000).

One of the components that diversify adaptive tests from linear tests is the item selection method. In this respect, it determines how the test will proceed from beginning to end, depending on the accuracy of the examinee's estimated ability (van der Linden & Pashley, 2009). There are many item selection methods in literature and many researchers are studied on the effectiveness of these methods based on both simulation and real practice. The source of motivation for proposing different methods are primarily the idea of obtaining more information with a different approach to more accurately estimate the ability, then the problem of exposure of the items over time and the fact that the items are composed of different contents (Thompson & Weiss, 2011). These methods are as follows; b-matching, (Urry, 1970), maximum Fisher Information (mostly used), Owen's Approximate Bayes Procedure (Owen, 1975), Maximum Likelihood Weighted Information (MLWI; Veerkamp & Berger, 1997), Maximum Global-Information Criterion (Chang & Ying, 1996), Maximum Posterior Weighted Information (MPWI; van der Linden,1998), Maximum Expected Information (MEI, van der Linden,1998), the progressive method (Revuelta & Ponsoda, 1998) and the proportional method (Segall, 2004).

The ability estimation method does not calculate only the final ability according to all responses of administered items but also interim ability which determines the next item. It also indirectly determines when the test is to be stopped. Therefore, it is one of the essential components of the CAT. The methods in the literature are maximum-likelihood estimator (MLE; Lord, 1980b) and weighted likelihood estimator (WLE; Warm, 1989); and two Bayesian Fashion method: maximum a posteriori (MAP; Samejima, 1969) and expected a posteriori (EAP; Bock & Mislevy, 1982). MLE is the most popular estimator in the past and uses the point at which the likelihood function is maximized to calculate the ability. Warm (1989) proposed WLE as an alternative to MLE in order to lessen the bias of MLE. Bayesian Fashion estimators use the posterior distributions of the ability but in different ways. For estimating ability, MAP as known Bayes Modal, maximize the distribution while EAP calculates its expected value (van der Linden & Glas, 2010).

The termination criterion deals with when and how to stop the adaptive administration of the item in the test. It is classified as fixed-length, ability precision level, the ability change, and the information level (minimum information). As with linear tests, the test is terminated after giving a predetermined number of items to the examinee in the fixed-length criterion. The ability precision criterion stopped the test when the accuracy of the examinee's interim ability is less than or equal to the predetermined value. The standard error of ability is used as a measure of accuracy. The criteria of ability change are stopped if the ability level changes to very small amounts after the items have been administrated. Finally, the information method finishes the administration of items if there is no available unused item that will cause a significant change in the information level. Also, combined versions of these criteria were used in different studies (Magis et al., 2017; Thompson & Weiss, 2011; van der Linden & Glas, 2010).

Online Calibration

Maintaining and extending the item bank by replacing exposed and outdated items with calibrated new ones are one of the fundamental requirements in ensuring the continuity of CAT. As mentioned in the introduction, the online calibration approach stands out among the other existing approaches to doing this. It is employed to calibrate new items (as denoted field test item or pretest items) which are randomly or adaptively administrated to examinees with operational items during the active testing by using their abilities (Stocking, 1988; Wainer & Mislevy, 1990). Examinees' responses to pretest items are not included in the scoring but are used only for calibration of the items. In this way, both the examinees’ abilities using operational items and the parameters of the pretest items using their abilities are estimated in CAT scenario.

Online calibration is associated with optimal design in terms of more efficient estimation of item parameters. Therefore, the statistical solutions in the optimal design have been applied to the online calibration procedure. This is implemented in two different ways; sampling the batch of examinee for pretest items and selecting appropriate items for examinees. Getting effective results in the studies (Berger, 1992; Berger, 1994) carried out by applying the optimal design to the paper-pencil test has been a driving force for its application to CAT (Zheng, 2014).

Online calibration is more effective and more prominent than traditional methods with its advantages. As mentioned above, one of the most important advantages is that it saves both money and time as it simultaneously estimates both ability and parameter (Makransky & Glas, 2010). Unlike the pretesting in the paper-pencil test, it is carried out during the operational testing, thus enabling the examinee to respond to new items with the same motivation by continuing the test mode (Parshall, 1998). Therefore, it gets more reliable data. Thanks to the parameter estimation methods for online calibration, it automatically places the pretest item parameters on the same scale as the operational items with no linking/scaling requiring additional operations (Chen et al., 2012). Since the assignment of each item to different samples reduces the rate of item exposure, the test security risk in online calibration is lower than the other approaches (Guo, 2016).

The online calibration procedure can be reviewed under the two following elements: pretest item selection method deals with the application of the pretest item and the estimation methods deal with how to calibrate the items. In addition, it also has the other elements deal with practical issues affecting the procedure. These elements are described below.

Pretest item selection methods. Pretest item selection method deal with the rules according to which pretest items are administered which examinee throughout the CAT session. It is one of the crucial elements that influence the pretest item calibration. This component has been handled by many researchers in different ways; optimal sequential design, random method and adaptive design.

Jones and Jin (1994), Y.C.I. Chang and Lu (2010), Lu (2014) and Zhu (2006) address this issue as a sequential or optimal sequential design. This design is based on the implementing of the traditional optimal design paradigm to the online calibration process (Zheng, 2014). The basic principle of the method is to select the suitable examinee for the pretest item from the examinee pool with different optimality criterion (L, E, A, mostly D). The static examinee pool is a requirement for this method. Despite that, this requirement cannot be met because CAT sessions are held at different times and each examinee who has completed the test leaves. Although this method is suitable for simulation, it is not feasible for these reasons in practice (Guo, 2016; Zheng, 2014)

In order to overcome the aforementioned problems, the proposed realistic way is to apply the most suitable pretest item for the calibration where the examinees reach the seeding position in accordance with the predetermined criteria. One of these criteria is the random selection method which is based on the randomly selection of the item from pretest item bank and very simple to implement (Wainer & Mislevy, 1990). However, since items are administrated in an adaptive way determined by the examinee’s ability in the CAT, the seeding of these randomly selected pretest items may distort this adaptive trend and cause these items to be perceived from operational items by the candidate. This may affect the examinee’s motivation during the exam in the negative way (He et al., 2020; Kingsbury, 2009; Zheng, 2014).

Another criterion is that the pretest items are administrated to the examinee in an adaptive way depending on the needs of the examinee or item rather than random way. Chen et al. (2012) and Kingsbury (2009) applied this method in line with the needs of the examinee as the adaptive nature of CAT. In this method, which corresponds to the called examinee-centered method by Zheng (2014), pretest items are selected in the same way as selection of operational items depending on the examinee’s ability. Although the different item selection methods in CAT literature could be used, Kingsbury (2009), Zheng (2014), and Zheng and Chang (2017) preferred MFI method. These methods aim to select the most appropriate item to maximize the accuracy of the ability parameter, not the accuracy of pretest items parameters. Therefore, adaptive design with examinee needs may not be competent (He, et al. 2020; Zheng, 2014). On the other hand, it showed efficient results for calibrating difficulty parameter in Zheng (2014) study. In addition, it had more accurate results than random item selection method (Kingsbury, 2009).

The adaptive method can also be applied in item needs form in which the comparing the examinee's contributions to the pretest items at the seeding position. The four different methods were applied depending on the adaptive design with item needs; Suitability index (SI; Ali & Chang, 2014), the comparison of D-optimal value, Ordered Informative Range Priority Index (OIRPI; Zheng, 2014), and Bayesian optimal design (van der Linden & Ren; 2015). SI method is calculated by considering the sample size and the target sample for ability ranges.

In the comparisons of D-optimal values methods, the pretest items are selected according to which of the item in the pool will have the maximum D-optimal value with the addition of the examinee's ability. The disadvantage of this method is that the items which have greater D-optimal value in the item bank tend to be administered much more than other items. This leads to ineffective parameter recovery for items with a lower D-optimal value (Zheng & Chang, 2017).

Zheng (2014) proposed the two different OIRPI method (OIRPI with Order Statistic and OIRPI with Standardization) based on the calculation of the D-optimal statistic at different ability ranges. In the OIRPI with Order Statistic algorithm, the examinee ability scale is divided into specific intervals. At the seeding position, the information of each item in these ranges is calculated. These information values are sorted for each range. After determining the range of the examinee's ability, the item that provides the highest information is selected by employing order statistic for this range. OIRPI with Standardization method is generally similar to Order Statistic, the information value is calculated by standardized to solve the problem that the information values are the same when the ability interval is small in the OIRPI with Order Statistic.

The Bayesian optimal design (van der Linden & Ren, 2015) calculates the expected contribution of the information to be obtained by adding the examinee’s ability to each pretest item at the seeding position. The critical keyword for this method is “the expected contribution” and means the information that may come from the assignment of the new item, which will be added to the information that the item has so far. van der Linden and Ren (2015) used D-, A-, E-, c- optimality criterion to obtain this information.

As seen in this section, D-optimality is widely preferred from the online calibration literature. This criterion is based on maximizing the determinant of the Fisher Information matrix calculated based on the item parameters. This means that the item is estimated with fewer measurement errors (Anderson, 1984).

Parameter estimation methods. The estimation method is concerned with how to estimate the pretest item in the online calibration procedure by making the use of operational items whose parameters are known. It is one of the most investigated issues in the literature along with item selection design. The methods differ in some points from traditional methods due to the used data structure. The reasons for the complicatedness of the online calibration are the fact that both operational and pretest items response are sparse because of the essential characteristics of CAT, calibrate as a basis of a restricted range of ability, and relatively small sample size from traditional paper-pencil administration (Stocking, 1988; Ban et al., 2001). In order to deal with these problems, several online calibration methods have been put forward for UCAT. These are Method-A/Stocking-A and Method-B/Stocking-B (Stocking, 1988), one expectation maximization (EM) cycle (OEM) method (Wainer & Mislevy, 1990), multiple expectation maximization (EM) cycle (MEM) method (Ban et al., 2001), BILOG/Strong Prior method (Ban et al., 2001), and maximum likelihood estimation-Lord’s bias-correction with iteration – Method A (MLE-LBCI-Method A) method (He et al., 2017). They are summarized as follows. As proposed by Ban et al. (2001), Zheng (2014) used Bayesian priors with Method-A, Method-B, OEM, and, MEM and tested the performances of these methods.

Method A. Among other calibration methods, Method-A (Stocking, 1988) has the simplest theoretical background and is easy to calculate (Chen & Wang, 2016). The basic principle underlying the method works as follows. First, the ability (θ) is calculated with maximum likelihood using operational item responses. Then, these are fixed and pretest item parameters are estimated using them by employing Conditional Maximum Likelihood Estimation process (Zheng, 2014). The pretest items parameter are on the same scale as the operational item because they are estimated the fixed thetas which are on the same scale as the operational items. Because of treating fixed abilities are as “true” ability without measurement error, it can cause parameter drift problems (Ban et al., 2001).

Method B. Although Method-B (Stocking, 1988) comes from the same basis of Method-A, it is strict and also impractical to implement. This is because it uses estimated anchor items to handle the parameter drift problem of Method A. In this method, candidates respond to operational items, pretest items, and anchor items, and then, the equating and transformation process is used to estimate parameters as similar Method-A (Ban et al., 2001).

One Expectation Maximization. The OEM method (Wainer & Mislevy, 1990) works primarily on calculating the expected posterior distribution of the ability estimated from administered operational items and then using it to maximize the Marginal Maximum Likelihood (MML) function. As a result of this method having only one EM cycle, the parameter is updated once with the distribution gained from the operational items only. In consequence of the distribution containing operational items, Parshall (1998) stated that the advantages of this method are that pretest items are on the same scale as the operational items and the calibration of any pretest item cannot infected by other pretest items

Multiple Expectation Maximization. The MEM method (Ban et al., 2001) is the derived form of OEM and the first EM cycle is operated in the same way in both. From the second EM cycle, the posterior distribution is calculated from both administrated operational and pretest items. Then, operational items parameters were treated as fixed and the process continues as in the OEM's maximization step. In this method, the pretest items also is same scale on operational items. It is an advantage of this method to utilize entirely the information get in the process, but it may be a disadvantage to include some deficiently estimated pretest parameters (Ban et al., 2001).

BILOG/Strong Prior Method. BILOG/Strong Prior method (Ban et al., 2001) performs the calibration of pretest using a computer program (Bilog-MG; Zimowski, Muraki, Mislevy & Bock, 1996) of the same name as the method. This method re-estimates the parameters of the operational items with strong prior distribution and then contribute them by fixing. Although there are similarities between this method and the MEM method in terms of employing the MML technique, the MEM method re-estimated pretest items parameters in cycles, not operational item parameters.

Maximum Likelihood Estimation-Lord's Bias-Correction with Iteration – Method A. MLE-LBCI Method A method (He et al., 2017) is proposed in order to eliminate the disadvantage of Method A from producing biased results in parameter estimation. For this purpose, Lord's bias-correction method (Lord, 1983) was implemented together with MLE in two different ways (at each update of the ability during the CAT test and at the end of the CAT test).

Other elements. Apart from pretest item selection methods and parameter estimation methods, the elements such as seeding position, termination rule, and sample size were discussed and their effects on item parameter estimation were examined in previous studies on online calibration.

The seeding position is defined as where the pretest item is to be assigned during the CAT session. Because the ability of before seeding is employed both in pretest item selection and parameter estimation, seeding position is one of the potential factors that can affect the calibration accuracy. Towards the end of the test in the adaptive session, theta includes fewer errors, thus providing maximum contribution to parameter recovery. The best option is to assign the pretest item very close to the end of the test. However, in this case, it could be a problem if the test takers notice it (Kingsbury, 2009; van der Linden & Ren, 2015).

There are many options for the seeding position in the literature. These are random assignment (Chen & Wang, 2016; Chen, 2017; Chen et al. 2017); fixed point (Kingsbury, 2009); at random position towards the end of the test (van der Linden & Ren, 2015); there location (early, middle, and late) of the test (He et al., 2020; Zheng, 2014; Zheng, 2016; Zheng & Chang, 2017). Each of these methods has different effects on parameter estimation and it is appeared that these effects can be significant or non-negligible in different studies (Chen & Wang, 2016).

The termination rule is the criterion that determines how to finish the sampling and when the pretest item export from the pretest item bank. This is another important factor to consider, as it is directly related to the accuracy of the parameter. While different criteria such as standard error and parameter stabilization have been recommended in the literature (Kingsbury, 2009), the sample size criterion is the most widespread in previous studies (Ali & Chang, 2014; Ban et al., 2001; He et al., 2017; He et al., 2020; Ren, van der Linden, & Diao, 2017; Kingsbury, 2009; Zheng, 2014; Zheng, 2016; Zheng & Chang, 2017; van der Linden & Ren, 2015). Ren et al. (2017) and van der Linden and Ren (2015) implemented a new criterion integrating of both posterior standard deviation and maximum sample size.

The sample size is applied in two different ways; total or cumulative sample size in all process and sample size for each pretest item. The first method is the number of candidates participating from the beginning to the end of the online calibration process, while the second method is the number of candidates required for the calibration of each pretest item. The pretest items may be exposed noticeably more than others in the total sample size method except when random selection is used. For instance, if D-optimal value comparison design or MFI criterion is used as pretest item selection method, some items could have greater D-optimal or MFI value than the others, therefore tend to selected primarily for all candidates. However, pretest items with the low value will be less selected and the calibration results will not be satisfactory relative to the others as they are not given to approximately the same number of candidates. Due to the second method focuses on each pretest, it may eliminate this problem of the first method.

Model, Concept, and Notations

The following concepts and notations are used to describe the issues in this study such as IRT model, pretest item selection methods and the parameter estimation methods.

Item Response Theory (IRT) is a model-based measurement theory that explains the relationship between an individual's response and his or her latent trait level such as ability, performance, intelligence, or competency level (Embretson & Reise, 2000). IRT uses a variety of statistical models describe the probability of correct response according to dichotomously or polytomously scored items.

IRT models defined with a mathematical function as called Item Response Function (IRF). Because of using dichotomous data, unidimensional dichotomous IRT model discussed in this study. These models range from simple to complex one-parameter logistic (1-PL) model (Lord & Novick, 1968; Rasch 1960), two-parameter logistic (2-PL) model (Birnbaum 1968), and three-parameter logistic (3-PL) model (Birnbaum 1968) respectively. Since the study was limited to the 1-PL model and 2-PL model, the notation of the 3-PL model was not included in this section. IRF for these models define the probability of a correct response $u_{ji} = 1$ of person $j = 1,2,...,N$ with ability θ_j on item $i = 1,2,...,L$ as;

1-PL

$$P_i(\theta_j) \equiv P_i(u_{ji} = 1 | \theta_j, b_i) = \frac{1}{1+exp[-(\theta_j - b_i)]}\ ; \qquad (1)$$

2-PL

$$P_i(\theta_j) \equiv P_i(u_{ji} = 1 | \theta_j, a_i, b_i) = \frac{1}{1+exp[-a_i(\theta_j - b_i)]}; \qquad (2)$$

where a_i and b_i are the discrimination and the difficulty parameters of item i, respectively.

Pretest item selection methods. This part describes the pretest selection methods which are examined the accuracy of parameter estimation in this study; MFI method, and D-optimal value design, and, Bayesian D-optimal design.

Suppose $k = 1,2,\ldots,m$ be the pretest item in pretest item bank and $\eta_k(a_k, b_k)$ be pretest item parameters which are need to estimate. Let $t-1$ items have already administrated pretest item of person j and R_t denote the eligible items in pretest item bank for tth item. In the online calibration design in this study, since the pretest items were administrated to the examinee after all operational items, θ_j that is the final ability of the examinee jth was used for the item selection. Suppose n_k be the number of examinees that have already responded item kth.

The methods used in this section are briefly explained by using statistical notation. The detailed explanations are presented the methodology section in Chapter 3.

Maximum Fisher Information. In online calibration context, this method aims to select the most informative pretest item among the remaining available items by using final ability. Therefore, the test information of the examinee is calculated using the item information function. For operational items $i = 1,2,\ldots,L$, the test item information is presented as

$$I(\theta_j) = \sum_{i=1}^{L} \frac{[P_i'(\theta_j)]^2}{P_i(\theta_j)(1-P_i(\theta_j))} \ , \qquad (3)$$

where $P_i'(\theta)$ is first derivative of the probability function $P_i(\theta)$ with respect to θ. With the use of test information function of examinee $I(\theta)$ given Equation 3 and the definition of $\arg\max$ that is an element or position or a point of functions where the function values are maximized according to a given argument, the most informative item k_t^* is selected as

$$k_t^* \ = \ \underset{k \in R_t}{\arg\max}\{I(\theta_j)\}, \ (4)$$

D-optimal value design. This design uses the D-optimality criterion that used Fisher information of the item k with η_k parameters by contributed θ_j. The information is defined this matrix. That is (Hambleton & Swaminathan,1985),

$$I(\eta_k; \theta_j) = \begin{bmatrix} I_{aajk} & I_{abjk} \\ I_{bajk} & I_{bbjk} \end{bmatrix}. \qquad (5)$$

Each element of $I(\eta_k; \theta_j)$ matrix is defined as;

$$I_{aajk} = (\theta_j - b_k)^2 P_k(\theta_j)(1 - P_k(\theta_j)) \qquad (6)$$

$$I_{abjk} = -a_k(\theta_j - b_k)^2 P_k(\theta_j)(1 - P_k(\theta_j)) \qquad (7)$$

$$I_{bbjk} = {a_k}^2 P_k(\theta_j)(1 - P_k(\theta_j)) \qquad (8)$$

Consider $\theta = (\theta_1, \theta_2, \ldots, \theta_{n_k})$ denote the ability vector of n_k examinee answering to item k and the total amount of information expressed as the sum of information provided by each ability. That is,

$$I_k(\eta_k; \theta) = \sum_{j=1}^{n_k} I(\eta_k; \theta_j) \qquad (9)$$

This method aims to maximizes determinant of total information matrix (denoted as D-optimal value/statistic). This also minimizes determinant of the covariance matrix and therefore reduces the measurement error of parameter estimation (Anderson, 1984). This method seeks k_t^* item that having the maximum D-optimal value using the ability of current examinee (denoted as θ_c) and θ by comparing all available pretest items. It is formulated as

$$k_t^* = \arg \max_{k \in R_t} \left\{ det\left[\sum_{j=1}^{n_k} I\,(\eta_k; \theta_j) + I(\eta_k; \theta_c) \right] \right\}. \qquad (10)$$

Bayesian-D optimality design. This method uses D-optimal design by modifying. It mainly focuses on the maximization of expected contribution with the θ_c. So that, it compares all eligible items and calculates their contribution (van der Linden, & Ren, 2015). It selects k_t^* using this equation given by;

$$k_t^* = \arg \max_{k \in S_t} \left\{ det\left[\sum_{j=1}^{n_k} I\,(\eta_k; \theta_j) + I(\eta_k; \theta_r) \right] - det\left[\sum_{j=1}^{n_k} I\,(\eta_k; \theta_j) \right] \right\}. \quad (11)$$

Parameter Estimation Methods. This part presented the parameter estimation methods for pretest items that are tested in this study; Joint Maximum Likelihood and One EM Cycle.

The likelihood function is used in parameter estimation. For ability θ_j, the likelihood of response u_{ij} to operational item i is

$$L\left(u_{ij}\middle|\theta_j\right) = P_i\left(\theta_j\right)^{u_{ij}}\left[1 - P_i\left(\theta_j\right)\right]^{1-u_{ij}} . \qquad (12)$$

The likelihood function of response v_{kj} to pretest item k for ability θ_j is expressed similarly,

$$L\left(v_{kj}\middle|\theta_j\right) = P_k\left(\theta_j\right)^{v_{kj}}\left[1 - P_k\left(\theta_j\right)\right]^{1-v_{kj}} . \qquad (13)$$

Joint Maximum Likelihood. JML is an estimator that simultaneously calibrates ability and item parameters. It is a two-stage iterative procedure. In the first step, the item parameters are treated as known and fixed, then the abilities are estimated. In the second step, it runs the opposite way round; the abilities that get in the first stage are treated as known and fixed, then the parameters are obtained (Baker & Kim, 2004; Hambleton & Swaminathan,1985). However, due to the final abilities of the examinees are known in this study, it is used only for the parameter estimation. Accordingly, this technique was applied separately to each pretest using its first stage. JML is defined using the log likelihood function. The joint likelihood for all response $v_k = (v_1, v_2, \ldots, v_{n_k})$ of n_k examinee to which pretest item k is administrated is the product of all separate θ_j's likelihood.

$$L = \prod_{J=1}^{n_k} L\left(v_{jk}\middle|\theta_j\right) = P_k\left(\theta_j\right)^{v_{jk}}\left[1 - P_k\left(\theta_j\right)\right]^{1-v_{jk}} . \qquad (14)$$

The principle underlying the parameter estimation of JML is to obtain $\hat{\eta}_k{}' = (\hat{a}_k, \hat{b}_k)'$ vector that maximize the log likelihood function. It means that the first derivative of the $ln\, L$ function with respect to $\hat{\eta}_k$ is zero. That is,

$$\frac{\partial \; ln\, L}{\partial \; \hat{\eta}_k} = 0 . \qquad (15)$$

This non-linear equation is solved by using the multivariate Newton-Raphson procedure since $\hat{\eta}_k$ column vector has two elements.

Suppose x be the $(d \times 1)$ column vector that maximize f function as $f(x)$. The general form of this procedure is presented as (Baker & Kim, 2004)

$$x_{t+1} = x_t - [f''(x_t)]^{-1} f'(x_t) . \qquad (16)$$

where $f''(x_t)$ $(d \times d)$ matrix and $f'(x_t)$ $(d \times 1)$ column vector are the second and the first order partial derivates of $f(x)$, respectively; and t is the iteration number. If it is applied to Equation 15, it is (Hambleton & Swaminathan,1985)

$$\begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix}_{t+1} - \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}^{-1} \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix}. \quad (17)$$

where

$$\ell_1 = \frac{ln\,L}{\partial \hat{a}_k} = \sum_{j=1}^{n_k} (\theta_j - \hat{b}_k)(v_{jk} - P_k(\theta_j)) \ , \quad (18)$$

$$\ell_2 = \frac{ln\,L}{\partial \hat{b}_k} = -\hat{a}_k \sum_{j=1}^{n_k} (v_{jk} - P_k(\theta_j)) \ , \quad (19)$$

$$\Lambda_{11} = \frac{ln\,L}{\partial \hat{a}_k^2} = \sum_{j=1}^{n_k} (\theta_j - \hat{b}_k)^2 (1 - P_k(\theta_j)) \left(\frac{v_{jk}}{P_k(\theta_j)} - P_k(\theta_j) \right) , \quad (20)$$

$$\Lambda_{22} = \frac{ln\,L}{\partial \hat{b}_k^2} = \hat{a}_k^2 \sum_{j=1}^{n_k} (1 - P_k(\theta_j)) \left(\frac{v_{jk}}{P_k(\theta_j)} - P_k(\theta_j) \right) , \quad (21)$$

$$\Lambda_{12} = \Lambda_{21} = \frac{ln\,L}{\partial \hat{a}_k \hat{b}_k} = -\sum_{j=1}^{n_k} \left(v_{jk} - P_k(\theta_j) \right) + \hat{a}_k \left(\theta_j - \hat{b}_k \right) \left(1 - \right.$$

$$\left. P_k(\theta_j) \right) \left(\frac{v_{jk}}{P_k(\theta_j)} - P_k(\theta_j) \right). \quad (22)$$

This iterative procedure is continued until it converges to the specified value or reaches the maximum number of iterations. After that, $\hat{\eta}_k(\hat{a}_k, \hat{b}_k)$ of the pretest item k are obtained.

One Expectation Maximization. OEM is a pretest parameter estimation method based on Marginal Maximum Likelihood Estimation (MMLE) method with EM algorithm similar to MEM. It uses the posterior distribution of abilities for more precise parameter estimation. It consists of two steps (E and M), and this EM cycle is executed once (Wainer & Mislevy, 1990).

In E step, the expected posterior log-likelihood of the pretest item k is obtained as the posterior ability distribution of n_k examinees that is given pretest item k. Suppose $n_i(i = 1,2,\ldots,n_i)$ be the number of operational items that an examinee reaches and $U_j = u_{ji} = (u_{j1}, u_{j2}, \dots, u_{jn_i})$ be the response of the ability j to these items. It also It is formed as the product of likelihood function from the responses of the n_k examinee to administrated operational items and their parameters $\hat{\eta}_{opr}$. That is,

$$L(U_j, \theta, \hat{\eta}_{opr}) = \prod_{i=1}^{n_i} P_i(\theta)^{u_{ji}} [1 - P_i(\theta)]^{1-u_{ji}} \ . \quad (23)$$

In M step, the posterior ability distribution is used to find $\hat{\eta}_k = (\hat{a}_k, \hat{b}_k)$ parameter that maximize marginal maximum likelihood (Wainer & Mislevy, 1990).

The EM cycle uses quadrature approximation approach to ensure continuity due to the integral-based definition of MMLE. Therefore, it uses quadrature points θ_h $(h = 1,2,\dots,q)$ on θ scale to a certain number and the weights of the distribution $W(\theta_h)$ corresponding to these points. Three different methods can be used in the selection of quadrate point; Gauss-Hermite quadrature point, quadrature over fixed points and Monte Carlo integration (Baker & Kim, 2004). For Gauss-Hermite quadrature methods, the weights are defined before. The method of quadrature over fixed points uses the density function of standard normal distribution (Mislevy, 1984). This parameter estimator also uses multivariate Newton-Raphson procedure. Before this procedure is executed, the expected value $\bar{n}_{kh}$ at each θ_h, and the expected number of correct responses $\bar{r}_{kh}$ at each θ_h are calculated as;

$$\bar{n}_{kh} = \sum_{j=1}^{n_k} \left[\frac{L(U_j, \theta_h, \hat{\eta}_{opr}) W(\theta_h)}{\sum_{h=1}^{q} L(U_j, \theta_h, \hat{\eta}_{opr}) W(\theta_h)} \right] ; \qquad (24)$$

$$\bar{r}_{kh} = \sum_{j=1}^{n_k} \left[\frac{v_{jk} L(U_j, \theta_h, \hat{\eta}_{opr}) W(\theta_h)}{\sum_{h=1}^{q} L(U_j, \theta_h, \hat{\eta}_{opr}) W(\theta_h)} \right] . \qquad (25)$$

The iterative Newton-Raphson algorithm is started with the implementation of the transformation $\hat{d}_k = log\ \hat{a}_k$. It is presented as

$$\begin{bmatrix} \hat{d}_k \\ \hat{b}_k \end{bmatrix}_{t+1} = \begin{bmatrix} \hat{d}_k \\ \hat{b}_k \end{bmatrix}_{t+1} - \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}^{-1} \begin{bmatrix} \ell_1 \\ \ell_2 \end{bmatrix} . \qquad (26)$$

where

$$\ell_1 = \sum_{h=1}^{q}[\bar{r}_{kh} - \bar{n}_{kh}P_k(\theta_h)]exp(\hat{d}_k)(\theta_h - \hat{b}_k)\ ; \quad (27)$$

$$\ell_2 = -\sum_{h=1}^{q}[\bar{r}_{kh} - \bar{n}_{kh}P_k(\theta_h)]exp(\hat{d}_k); \quad (28)$$

$$\Lambda_{11} = \sum_{h=1}^{q}\bar{n}_{kh}[exp(\hat{d}_k)]^2(\theta_h - \hat{b}_k)^2\, P_k(\theta_h)(1 - P_k(\theta_h))\ ; \quad (29)$$

$$\Lambda_{22} = -\sum_{h=1}^{q}\bar{n}_{kh}[exp(\hat{d}_k)]^2\, P_k(\theta_h)(1 - P_k(\theta_h))\ ; \quad (30)$$

$$\Lambda_{12} = \Lambda_{21} = -\sum_{h=1}^{q}\bar{n}_{kh}[exp(\hat{d}_k)]^2(\theta_h - \hat{b}_k)\, P_k(\theta_h)(1 - P_k(\theta_h)). \quad (31)$$

This procedure is continued until it converges to the specified value or reaches the maximum number of iterations. After that, the final transformation $\hat{a}_k = exp(\hat{d}_k)$ is employed and $\hat{\eta}_k(\hat{a}_k, \hat{b}_k)$ of the pretest item k are obtained.

Previous Research

In this section, after the classification of previous studies, each of these is summarized. The online calibration procedure is extended and integrated into all CAT design; Unidimensional CAT (UCAT), multidimensional CAT (MCAT). Since this study is based on IRT, UCAT studies and MCAT studies were involved in this section while Cognitive Diagnostic CAT (CD-CAT) studies were not included.

Previous Research About Unidimensional CAT. In this section, the studies about unidimensional CAT are listed as follows; Buyske (1998), Chang and Lu (2010), Lu (2014), Makransky and Glas (2010), Ban et al. (2001), Kingsbury (2009), Ali and Chang (2014), Zheng (2014), van der Linden and Ren (2015), He (2015), Zheng and Chang (2017), He et al. (2017), He et al. (2020) and Zheng (2016).

Buyske (1998) proposed the L-optimal design as an alternative to the D-optimal design. It is based on the two-point design that the probability of responding correctly to the item is equal weight to 25% or 75%. In the study, standard (random), D-optimal, and proposed (L-optimal) design were compared on parameter accuracy using MSE, bias and sample size criteria. According to the findings, the standard methods have a larger error than the proposed methods chosen for extreme parameter b values. In general, the methods except standard method have similar results in terms of MSE and bias. On the other hand, L-optimal design needed smaller sample size than L-optimal design in similar condition and it was lightly more effective and practical in this respect.

Chang and Lu (2010) discussed online calibration in the context of sequential design under variable length CAT. Their designs consist of two stages CAT and item calibration process. In the first stage, examinees abilities are estimated with conducted CAT procedure. In the second stage, the pretest items are calibrated by selecting the most informative examinee using the 2-point D-optimal design. In addition, the measurement error models are applied to compare the parameter estimation accuracy. Two studies were performed by using simulated and real data in 2-PL. The results showed that the sequential design performed well at the parameter estimation and this method was more effective than the sample size for both data. In addition, they stated that the method could be used in large scale examinations.

Lu (2014) studied the effect of different sequential optimal design methods on parameter estimation. In the study, D-optimality, A-optimality, E-optimality, and random design have been tested in low to high discrimination (0.5 to 2.5) and difficulty parameter (-3 to +3) conditions under the 2-PL. MSE and sample size range-based stopping time were used as a criterion to compare the results. There is no significant difference between the compared methods when estimating *a* parameter. For the *b* parameter estimation, the most advantageous methods in parameter accuracy are E-optimal and A-optimal when the parameter a is low. On the other hand, while the parameter a is high, D-optimal and A-optimal are the most effective methods. In general, optimal methods were found to be more appropriate than random methods. The sample size increased when parameter a reached from low to high. A similar trend is also seen as the parameter b is close to -3 and +3. In addition, the sample size in the random method is lower than the optimal method because examinees are not selected according to the suitability of the parameters.

Makransky and Glas (2010) explore new adaptive model to make more practical and rapid calibration that can be used in small testing. For this purpose, they tested three different optimal calibration methods with a simulation study. These strategies are two-phase, multi-phase and continuous updating strategy. These strategies are based on the random or adaptive selection of the items in stages and vary according to the number of steps. In the study, the transition points

(10, 25, 50, 100 and 200 item administration), the sample size (250, 500, 1000, 2000, 3000 and 4000), the Item Response Theory (IRT) model (1-PL and 2-PL) and the item pool (200 and 400) conditions are simulated. The continuous updating strategy had better results than the other strategies in all stage points and item pool size. The weak point of this method is that it requires more application for parameter estimation of a few items in the 2-PL. Multi-phase strategy showed good results in large samples in the 2-PL model. In addition, the two-phase strategy generally has lower performance than the others.

Ban et al. (2001) examined the effect of five pretest item calibration method (Stocking-A, Stocking-B, OEM, MEM, and BILOG with Strong Prior) on pretest item recovery in online calibration. For this purpose, the performance of these methods was investigated under different sample size (300, 1000 and 3000) with a simulation study. This study used 540 items in total for operational item bank, pretest items and linking items respectively 520, 10, and 10. The adaptive tests were applied on fixed length (30 items). The item characteristic curves (based on weighted mean squared error-WMSE), bias, standard error (SE) and root mean square error (RMSEA) was used as a criterion for evaluating the performance of pretest item calibration methods. The findings of the study show that the method with the smallest parameter estimation error in all sample sizes is the MEM. However, it has the most time-consuming method because its procedure is iterative. In spite of needing the anchor items and big sample size, the Stocking’s Method B is the second-best option for pretest item calibration. The OEM has larger error than MEM. The Stocking’s Method A is the last option because of having the largest parameter estimation. The BILOG with Strong Prior method requires a big sample size so that it is not feasible for small sample size. On the other hand, this method is suitable for sparse data.

Kingsbury (2009) explained how the online calibration process works as an adaptive item calibration with an example and discussed the specific components that need to be controlled in the process. In addition, an online calibration simulation study was performed depending on sample size and item selection in the study. Random sampling and adaptive item calibration methods were used as item selection methods and were compared. The study was limited to the 1-PL model and online calibration was continued until the number of responses has extended

500 (as also sample size). As the evaluation criterion, the mean absolute parameter difference, correlation coefficient and bias were calculated. The results show that in the all samples sizes less than 150, the adaptive item calibration showed fewer errors than the random sample method. For two methods, similar results were found in the case of large samples size greater than 150. In terms of correlation coefficients, it was seen that the adaptive item calibration obtained a larger correlation coefficient in all sample sizes. In terms of bias, the random sample item calibration method with 40 response showed the greatest positive bias, while adaptive item calibration with 10 response showed the greatest negative bias. However, as the sample size increased, the bias also decreased for both methods.

Ali and Chang (2014) introduced the Suitability Index (SI) as a new pretest item selection method developed by them and tested the effectiveness of this method according to other methods. The effects of the sample size (300, 500, 750, 1000, 1500 and 2000) were tested in the Monte Carlo simulation study. The research mainly focuses on the effect of the pretest item selection method. These methods are the Maximum Suitability Index (MSI), match-b and random. The performance of the methods in the study was examined by calculating the root mean square error (RMSEA) and bias for pretest items parameters. For discrimination and difficulty parameters, MSI showed fewer measurement errors than match-b and random method. MSI and match-b methods' performance is relatively better for guess parameter, but the results are very close to each other. In terms of sample size, bias is low for all item selection methods in small samples size and RMSE decreased as the number of samples increased for the discriminant parameter. Both bias and RMSE became smaller in all methods as the number of samples increased for difficulty parameters. For guess parameter, the results are very similar all sample size.

Zheng (2014) introduced the Order Statistics (OIRPI-O) and Standardization (OIRPI-S) methods based on The Ordered Informative Range Priority Index (OIRPI) for the selection of pretest items for the replenishment of the item pool. In addition, the performance of these methods was compared with other pretest item selection methods (Random, Examinee Centered, Optimal-D). The parameter estimation accuracy of the seeding location (early, middle and late), pretest item calibration methods (Stocking-A, OEM, MEM, Bayes Stocking-A, Bayes OEM and Bayes MEM) was examined for 1-, 2-, and 3-PL model in the study. The mean square of the error squares (RMSE) and the difference between the estimated and real item characteristic curves were calculated to compare performance in the study. Findings of the study indicated that both methods of OIRPI were the most effective in all IRT model and item parameter. In the 2- and 3-PL model, the examinee centered method showed more accurate results for the *b* parameter whereas it showed worse results in the other parameter accuracy. In terms of estimation methods, Bayesian MEM showed more balanced results than other estimation methods. When the effect of the seeding location for OIRPI-O and OIRPI-S methods are examined, it is found that the middle and late seeding more accurate estimations of item parameter than the early seeding. The increasing trend was shown from early seeding to late seeding for a and c parameter in the examinee centered method. Middle and late seeding results were similar in the other methods.

van der Linden and Ren (2015) suggested an adaptive estimation design using the Bayesian criteria for the field test item. The design uses both the posterior distribution of the examinee’s abilities acquired as adaptive and the posterior distribution of the field test item parameters applied consecutively. Two simulation studies were performed for two different purposes. The first of these was to analyse of the two MCMC applications in terms of parameter accuracy. The second of these was to test the performance of Bayesian D-optimality, A-optimality, c- optimality for only parameter a and random item selection criteria. In the first study, nine item parameters were estimated by the MCMC algorithm with 12000 iterations. In the findings, it was found to be effective on parameter precision with small variance for difficulty and guessing parameter (unexpected case for it). In contrast, less accurate results were obtained for the discriminant parameter. In the second study, 50 field test items were compared in terms of completing the item calibration according to four different items selection design and two stopping rules (below the threshold for the item’s posterior standard deviation and predetermined sample size). In standard deviation stopping rule, the D-optimal method had the tendency to retire earlier by calibrating items from A-optimality and random design except for last a few items. The random method is more efficient than A-optimality. In fixed sample size rule, D-optimal showed similar results to the previous results for all items. In terms of performance, A-optimality and c-optimality for a-parameter showed similar results while the random method had the worst performance.

He (2015) examined the optimum sample size for pretest item (as defined field test item) calibration under Rasch Model. In this study, the sample size was used as the number of responses required for each pretest item. Different sample size (30, 60, 120, 250, 500 and 1000) was compared with Monte Carlo simulation study. Bias, absolute bias (abias) and mean squared error (MSE) was used to measure the pretest item calibration accuracy and precision. According to the results, difficult items were overestimated, while easy items were underestimated. As the number of samples increased, the bias, abias, and MSE were decreased. When the sample size was 250, the parameter recovery was acceptable and sufficient. However, when the sample size reached to 1000, almost unbiased results were obtained for parameter precision. Another result is that the contribution to the parameter estimation of the bias correction formula was slight.

Zheng and Chang (2017) aimed to compare five different pretest item selection methods with a simulation study. The methods discussed in the study are random selection, the examinee-centered method, D-optimal method, Bayesian D-optimal design, and The Ordered Informative Range Priority Index (OIRPI). In this study, the examinee-centered method is applied as Maximum Fisher Information selection method. Parameter estimation methods (OEM and MEM), seeding location (early, middle and late) and sample size (1000, 1500, 2500, 5000 and 7500) are other elements examined in the study. The study also simulated under conditions of the 1-PL, 2-PL, and 3-PL. In order to examine the effectiveness of the methods in different situations, pretest item parameters were formed by crossing parameters *a* and *b* divided into different levels from low to high. In terms of estimation methods, the results show that OEM is as accurate as MEM in many conditions. On the other hand, OEM produced abnormal results when examinee-centered method was used in estimating large *a* parameter in the 2-PL and 3-PL. In general, OEM is suitable for 1-PL, while MEM is more suitable for 2-PL and 3-PL as it produces more consistent results. The seeding location effect was not particularly noticeable especially in the 1-PL model and most cases in the 2-PL and 3-PL. As expected, more accurate estimation results were obtained in the middle and late seeding locations for examinee-centered and ORPI methods. The increase in the sample size gets better the accuracy of parameter estimation and this change occurred at various rates especially in the 2-PL and 3-PL depending on the high and low values of *a* and *b* parameters. On the other hand, the accuracy of the c parameter estimation gets worse with increasing sample size. In terms of estimation methods, findings are different from previous studies. Although OIRPI and examinee-centered method improve the parameter accuracy for 1-PL, the difference between other methods is quite small. In the 2-PL model, it was found that adaptive methods are more effective than the random method only under some conditions (high *a* and small absolute *b* parameter). The interesting result is that the random method, which is a simple method, gives similar results and also to be more effective than many complex adaptive methods.

He et al. (2017) suggested a new online calibration method to improve the performance of the Method-A. In this method, named MLE-LBCI-Method A, the ability estimation is performed the corrected MLE (Lord's bias-correction method) and Method-A is used in the online calibration part. Two simulation study were performed in which the corrective MLE in CAT settings was examined and the performance of MLE-LBCI-Method A was compared with the other methods (Method A, OEM, and MEM). In addition that, sample size (1000, 2000, and 3000) and test lenght (10, 20, and 30) was examined as a simulation condition. RMSE, bias and the area difference of item characteristic curves (AWG) were used as evaluation criteria. The new method showed better results than Method A in most cases. It also showed less AWG value than OEM in some condition. In terms of RMSE, it had less error than MEM on calibrating a and c parameter in most case. MEM is the best method in all conditions but it needed more time than others. As the sample size increases, the precision of the item parameter estimates increases.

He et al. (2020) aimed to achieve the effectiveness of calibration in online calibration by suggesting the excellence degree (ED) criterion by integrating the original D optimal design with modified d-optimality design (called in D-VR design) introduced by van der Linden and Ren (2015). Four different design schemes (original estimated information (o), the minimum information (min), the mean information (mean), and the likelihood-weighted information (lw)) based on the ED criteria were created and compared with the DV-R design. Three different simulation studies were conducted. In the first study, the effect of the number of pretest sample size (200,400, and 600) and the seeding positions (early, middle and late) were examined. In the second study, the effect of sample size at random calibration stage and the number of parameter update samples of the pretest items were tested. Finally, the accuracy of calibrated items in ability estimation was investigated. In the findings, all ED designs show more improvements over the DV-R design under the most situation. It was found that parameter estimation was more effective in cases when new items were seeded in the middle and last position in the test. ED-o produced more accurate results than D-VR design in the second simulation. The ED-o design is efficient when the sample size at random stage is fairly small, and vice versa for the DV-R. The increase in the number of parameter updates provided only time-saving efficiency but did not cause much change in calibration accuracy. As predicted, for the accuracy of the ability parameter the calibrated items had a worse result than the actual ability parameters in the last simulation study. Whereas D-VR worked better than ED-o design when the sample size is small, ED-o results are more accurate when the sample size increases. In addition, the most prominent features in schemes except than ED-o were the following. ED-lw was the best solution when seeding position was early. Another effective design when the pretest item is seeded early was ED-min but its process takes a long time. It was found that ED-mean performs quite well when the small standard error range was used and the seeding position was middle and late.

Zheng (2016) explained how the online calibration process works under the generalized partial credit model for polytomous items and then examined the accuracy of parameter estimation methods with a simulation study. For this purpose, the effect of estimation method (OEM and MEM), number of response category (3 and 4), pretest item selection method (random, maximum θ information and match-b), seeding location (early, middle and late) and calibration sample size (200, 500, and 500) were investigated as a simulation condition. The parameter estimation accuracy in different conditions was examined by calculating RMSE for each parameter. The results showed that in the OEM method, the RMSE rises when the number of categories increases from 3 to 4. On the other hand, there is no difference in the number of these two categories when using MEM. From the point of view of pretest item selection methods, it has been found that the random method had more accurate than maximum θ information and match-b in many conditions. In terms of seeding location. the results are similar. The results show that the RMSE based on sample size have changed according to the method of pretest item selection and calibration method and when the sample size is 4000, the results are not different from each other.

Previous Research About Multidimensional CAT. In this section, the studies about multidimensional CAT are listed as follows; Chen and Wang (2016), Chen et al. (2017) and Chen (2017).

Chen and Wang (2016) argued the fact that when the M-Method A method is used for parameter estimation in the multidimensional CAT, the ability is considered as real ability and therefore is not to take any notice measurement errors of its. In order to overcome this deficiency, they have proposed two new parameter estimation methods (FFMLE-M Method A – Individual and Mean) by combining with the M-Method A and the full Functional MLE. The effect of these method and other methods (M-Method A and M-MEM) on parameter recovery was tested under simulation conditions based on item pool type (within-item and between-item design), sample size (1500 and 3000) and test length (20 and 40). The results show that the two proposed methods had better performance than Method-A on parameter accuracy, especially discrimination parameter, in all large sample size conditions. This performance is more noticeable in short tests. In contrast, when the sample is small, these two methods are not effective in estimating the parameters as much as the original M-Method-A. In addition, these methods also performed better than M-MEM.

Chen et al. (2017) developed new and favorable online calibration methods (M-Method-A, M-OEM, and M-MEM) for multidimensional CAT (MCAT) and tested their effectiveness for item parameter recovery. These methods are based on the methods used in the unidimensional CAT. NAMC and AMC integration is used to adapt these methods (M-OEM and M-MEM) to MCAT. Besides the methods, item bank design (within and between item), the correlation between dimensions (.0, .5 and.8) and test length (20,30,40 item) were also examined. In the scope of the study, three simulations were conducted, two of which were based on simulated item bank and one was based on a real item data. In the first simulation study, the effects of NAMC and AMC integration were examined in terms of parameter recovery (ability and especially item). The variables tested in the second simulation were the effect of random item selection and the proportion of sample size (%25, %50, %75) in terms of pre-calibration sample size (random item application). In the last simulation study, the conditions in the study were investigated by using real item parameters. According to the results obtained from the first simulation study, three developed methods show similar results and provide precise item parameter estimation. In five of the six tested conditions, M-OEM showed the best results, while in the remaining one, M-MEM had the best results. The accuracy of the ability and item parameter estimation is directly proportional to the increase in test length. Another result obtained is that NAMC and AMC integration give similar results. In addition, the proposed AMC integration was found to be more useful in estimating parameters in difficult situations. In the second simulation, adaptive calibration design has more advantages than random calibration design in most conditions. The relationship between the changes in the proportion of sample size could not be obtained. The third simulation results showed that abilities for both dimensions can be accurately estimated. Unlike the second simulation condition, increasing the proportion of sample size in the adaptive design estimate parameters with fewer errors in the proposed three online calibration methods.

Chen (2017) put forward two new estimators (M-OEM-BME and M-MEM-BME) in the multidimensional computerized adaptive test (MCAT) literature by integrating Bayesian model estimation with multidimensional OEM (M-OEM) and multidimensional MEM (M-MEM) methods. These two methods use the information of a priori distribution both the ability and the items' parameters. The two simulation studies were conducted under 9 conditions formed by crossing 3 level dimensions of ability (no correlation, moderate and strong) and 3 level sample size (900, 1800, 3600). In the first simulation study, existing six (M-Method A (True), M-Method A (Original), FFMLE-M-Method A (Mean), FFMLE-M-Method A (Individual), M-OEM, M-MEM) and the new two parameter estimation methods were compared using random design in terms of ability and item parameter recovery, the number of EM cycle and time consumption. The results of the first simulation showed that: M-MEM-BME had the best results in parameter estimation in cases when there is no correlation between the dimensions; FFMLE-M-Method A (Individual) was effective in large samples, and M-OEM OEM had satisfactory performance results under the other conditions. Besides, M-MEM-BME found a solution to the non-convergence problem of M-MEM. However, both methods are not particularly economical in some cases (strong correlation between dimensions and large sample sizes) in terms of time consumption. In the second simulation, the random design and the optimal Bayesian design proposed by van der Linden and Ren (2015) compared using FFMLE-M-Method A (Individual) estimator. It was found that the difference between methods on the parameter recovery was too close to be considered.

Summary of Previous Research. In the literature, some of the calibration studies (Buyske,1998; Chang & Lu, 2010; Lu, 2014) used the principle of examinee selection for the item from examinee pool, in other words, sequential design. The preferred designs in these studies are L-optimal, A-optimal, E-optimal and mostly D-optimal. Makransky and Glas (2010), unlike others, proposed a new design called automatic online calibration for small testing programs. On the other hand, the most preferred design in the literature is adaptive design. The specified version of this design, the Bayesian optimal design proposed by van der Linden and Ren (2015), has been a source of motivation for other studies (Chen, 2017; He et al., 2020; Zheng & Chang, 2017). Among the optimality methods, D-optimality is frequently employed in these designs as a result of its structure and success in calibration. In addition, the random design is widely compared with them in terms of performance. Although quite simple, it performed similarly impressive with other complex designs (Chen, 2017; Zheng & Chang, 2017)

Apart from the design, another factor that is considered in the previous studies is parameter estimation method. These methods were proposed primarily for UCAT and then generalized for MCAT and CD-CAT; Stocking-A, Stocking-B, BILOG, OEM, and MEM. The simplest method, Method A, was performed in extended version by combining Lord bias correction in UCAT (He et al., 2017) and full functional MLE method in MCAT (Chen & Wang, 2016; Chen, 2017). The other two important and frequently used methods are MMLE with OEM and MMLE with MEM. They were modified with Bayesian prior to get better performance. The expected effect on the parameter recovery had seen in Zheng (2014) and Cheng (2017) with this modification. It has been found that MEM was better in most cases in parameter estimation due to the fact that having multiple cycles in UCAT and MCAT versions. However, a consequence of this causes a loss of time. Besides, OEM has worked in some cases as effective as MEM or even more effective (Chen et al., 2017; Chen, 2017; Zheng & Chang, 2017).

Finally, the investigated factors are the sample size as termination rule for online calibration procedure and the seeding position for pretest item. As expected, the number of sample size is directly proportional to the accuracy of pretest item parameter estimation (Ali & Chang, 2014; Chen et al., 2017; He, 2015; He et al., 2017; Kingsbury, 2009; Zheng & Chang, 2017). On the other hand, in Zheng and Chang (2017) study, this trend continued for parameters *a* and *b*, and surprisingly, with the increase in the sample size, the parameters c was estimated less precisely. Due to having less error estimation of ability towards end of test, it was found that middle and late seeding were more effective than early seeding (He et al. 2020; Zheng, 2014; Zheng & Chang, 2017).

Based on these aforementioned studies, in this study, the item selection method (MFI, D-optimal value design, and Bayesian-D optimality design), parameter estimation method (JML and OEM), sample size of both random stage (250, 500, and 1000) and all calibration process (250, 500, and 1000) components which have effect on parameter estimation (as mentioned in the studies in the online calibration literature above) were considered as variables together and their effect on parameter estimation accuracy and cumulative sample sizes were examined with a simulation study.

Chapter 3
Methodology

Type of Research

In this research, the effect of the pretest item selection methods, parameter estimation methods, the sample size of random calibration stage, and the calibration sample size of pretest item on the accuracy of the item parameter estimates and cumulative sample size were investigated in the online calibration scenario. The data (operational and pretest item parameters and examinees' ability parameters) in this research have been generated and the online calibration process has been simulated with a computer program. As with any simulation study, this study was carried out under certain conditions that restrict the generalization of the results (Davey, Nering, & Thompson, 1997; Feinberg & Rubright, 2016). For all these reasons, it can be classified as a Monte Carlo simulation study. These types of studies are commonly used in psychometry and make it possible to evaluate and compare the performance of different methods (Rubinstein & Kroese, 2017).

Simulation Study Design

An online calibration procedure was carried out with the simulation study for the purpose of the study. The simulation study was conducted using 'Rcpp' package (Eddelbuettel et al., 2018) in R (R Core Team, 2017) with the computer program which is written by the researcher. The features and development of this program are detailed in the Online Calibration Computer Program section. The calibration process was carried out based on 1-PL and 2-PL IRT models, separately. Since the items are targeted or tailored for the examinees, there is not much guessing involved in a CAT administration (Glas, personal communication). Therefore, 3-PL IRT model was not considered in this study.

In this study, an online calibration procedure was applied in two phases. In the first phase, pretest items were administered to an examinee at the end of the administration of operational items within his/her individual CAT-session. Following that, responses and ability estimates were recorded. At this phase, pretest items were administered randomly. This phase continues until the number of observations/examinees reach a certain number (defined as the sample size of the random calibration stage). Due to the fact that enough observation was not obtained for pretest items within this phase and therefore parameter estimation was unstable, parameter estimation was not performed; hence this stage is also called the pre-calibration phase. At the end of the first phase, pretest items have a minimum number of observations that allow their parameters to be estimated. In this way, the pre-calibration phase is completed and item parameters were estimated separately conditional on the examinees' ability estimates and responses to operational and pretest items. Before the second phase, each pretest item has a temporary initial parameter value. In the second phase, given the item parameter estimates of the first phase, and the ability estimate at the end of the CAT, pretest items were chosen by using all available information on item parameters, ability parameters, and response pattern according to applied item selection method. During this process, the administration of pretest items to examinee continues and the pretest items' parameters were updated periodically depending on the number of responses. In this process, once a pretest item reached the predetermined number of responses (defined as the calibration sample size of pretest item), the calibration process for that item ended and it was removed from the pretest item bank. The process continued until no pretest item remained in the pretest item bank (The number of examinees required for all process is also defined the cumulative sample size.) Throughout this online calibration process, the adaptive test was administered to examinees one after another. The responses of examinees to the pretest items in two phases were not included in the scoring.

In this study, IRT models (1-PL and 2-PL), pretest item selection methods (MFI, D-optimal value design, and Bayesian D-optimal design), parameter estimation methods (JML and OEM), the sample size of random calibration stage (250, 500, and 1000), and the total number of responses (250, 500, and, 1000) were the factors that were simulated across the conditions to examine the effects of these on the parameter estimation in accordance with the simulation design described above. For each condition examined in the study, the simulation was repeated 100 times (i.e. *rep* = 100) to reduce random errors. Simulation conditions are summarized in Table 1.

Table 1

Simulation Conditions

Simulated Factors	Methods	Number of Conditions
IRT model	1-PL 2-PL	2
Sample Size of Random Calibration Phase	250 500 1000	3
Pretest Item Selection Methods	MFI D-optimal value design Bayesian-D optimality design	3
Pretest Item Parameter Estimation Methods	JML OEM	2
Calibration Sample Size of Pretest Item	250 500 1000	3

Generation of Item Parameters and Examinees

For each replication, the operational item bank containing 250 items and the pretest bank containing 25 new items (i.e. m = 25) were randomly generated from the following the distribution. As in the simulation design of He et al. (2020), Kingsbury (2009), and Zheng (2014), different operational and pretest item banks were used for each replication in this study. The parameters *a* were drawn from log-normal (0, 0.25) distribution and the parameters *b* were drawn from standard normal (0, 1) distribution. Similarly, these distributions were used in previous studies (Fink, Born, Spoden, & Frey, 2018; Natesan, Nandakumar, Minka, & Rubright, 2016). The reason for selecting these distributions is to resemble real situations for the items. The descriptive statistics of all items in the different operational and pretest item banks are presented in Table 2. As with the generation of the item parameters, the ability parameters for each test taker in each replication were randomly sampled from standard normal (0, 1) distribution.

Table 2

Descriptive Statistics of Operational and Pretest Item Bank

Statistics	Operational Item Bank			Pretest Item Bank		
	1-PL	2-PL		1-PL	2-PL	
	b	*a*	*b*	*b*	*a*	*b*
Mean	-0.051	1.016	-0.051	-0.002	1.025	-0.002
Std. Dev.	0.992	0.241	0.992	0.996	0.261	0.996
Min	-2.467	0.519	-2.467	-3.162	0.453	-3.162
Max	3.100	1.980	3.100	3.108	2.256	3.108

Adaptive Test and Online Calibration Procedure

In this section, the details of the adaptive testing process and the online calibration procedure are explained, and then the item selection methods in the research are introduced.

As mentioned in the simulation, firstly the operational items administrated to the examinees. The medium items with difficulty parameters ranging between -0.5 and 0.5 were selected randomly as the initial operational item as proposed by Thompson and Weiss (2011). MFI criterion was applied for the selection of the following item. Both interim and final ability were estimated by using weighted maximum likelihood estimator (WLE; Warm, 1989). An examinee responds to 35 operational items before reaching the pretest items. To put it another way, the fixed number of 35 items was used as the termination rule for the application part of the operational items in CAT session.

After administration of operational items, pretest items were started to be seeded to the examinees. In this study, the seeding position was determined as the end of operational items. The motivation for this is that the final ability estimated at the end of the test include least measurement errors and greatest information (van der Linden & Ren, 2015), and as a consequence, the use of these abilities for item parameter estimation provides more accurate results. The number of pretest items applied to each examinee was determined as 5 or less (i.e. $D \leq 5$). Towards the end of the online calibration process, fewer than 5 items are administered since there is not enough pretest items in the item bank whose calibration process is not finished.

The pre-calibration phase continued until the responses to the pretest items reached a certain number, in other words, the sample size of the random calibration stage. Due to the number of observations had an effect on the accuracy of parameter estimation (Chen and Wang, 2016), the three levels of the sample size were compared for random calibration stage: 250, 500, and 1000 (i.e. n_r = 250, 500, 1000). Because the pretest items are randomly assigned to the examinees in this phase, the number of responses that the average for each item has to calibrate is equal to the product of the sample size of random calibration stage and the number of pretest items applied to each examinee divided by the number of items in pretest item bank to be calibrated [i.e., $(n_r \times D)\ /m$]. Given these sample size of random calibration, each item has an average of 50 [i.e., ((250 x 5)/25)], 100 [i.e., ((500 x 5)/25)], and 200 [i.e., ((1000 x 5)/25)] responses, respectively. At the end of the first phase, the initial parameters of the pretest items were calculated using abilities and responses.

In the second phase of the calibration procedure, when an examinee reaches the seeding position, the pretest item is selected according to the applied item selection method and administrated. In this study, the three pretest item selection methods were employed: Maximum Fisher Information method, and D-optimal value design, and Bayesian D-optimal design.

During the second phase, the pretest items parameter was updated as each item reached each new sample. The parameter update sample size is set to 10 new responses which is one of the lowest numbers used in previous studies (He et al., 2020; Zheng, 2014; Zheng & Chang, 2017). The same estimation technique was used in the computation of the initial, interim and final parameters of pretest item parameter. Depending on the simulation conditions, the performance of two pretest item parameter estimation method was examined; Joint Maximum Likelihood and One EM Cycle.

For both JML and OEM, parameter estimation procedure is iterated until the absolute change in estimated values was less than the threshold value or reaches the maximum number of iterations. The convergent threshold and the maximum number of iterations were set 0.001 and 100, respectively. The online calibration procedure is terminated when each item reaches a specified fixed number of responses/examinees. The three-level calibration sample size of pretest item were investigated; 250, 500, and 1000 (i.e. N = 250, 500, 1000). The reason for selecting these sample sizes is that 250 (He, 2015; He et al., 2017), 500 (He, 2015; He et al., 2017; He et al., 2020; Kingsbury, 2009; Zheng, 2014; Zheng, 2016) and 1000 (He, 2015; van der Linden & Ren, 2015; Ren et al., 2019; Zheng, 2016) are widely preferred in the previous studies.

Pretest item selection methods. In this part, Maximum Fisher Information method, and D-optimal value design, and Bayesian D-optimal design are introduced.

Maximum Fisher Information. MFI method is a standard item selection method commonly used in CAT. This approach aims to select the operational item that maximizes Fisher information at interim theta based on the test items previously administrated in the exam (van der Linden & Pashley, 2000; Weiss, 1982). Although this method is intended to optimize the ability estimation, it can optimize the accuracy of parameter estimation, especially in the 1-PL model by matching b parameter to ability during the online calibration process (Zheng, 2014). In addition, it was used in different studies and it was found to be effective compared to other item selection methods in some conditions. For these reasons, this criterion is included in this study as a pretest item selection method (Kingsbury, 2009; Zheng, 2014; Zheng & Chang, 2017).

D-optimal value design. The D-optimal value design is mainly related to optimal design. The basic principle behind this design is to minimize the standard error of parameter estimation by maximizing the determinant of the Fisher information matrix (Anderson, 1984). This design was applied in some studies in the online calibration literature (Buyske, 1998; Chang & Lu, 2010; Jones & Jin, 1994; Zhu, 2006) to select the appropriate examinee from the examinee pool for the pretest item. As mentioned previously, it was thought that this would not be suitable because of the continuous nature of the CAT sessions. Therefore, as in Guo (2016), Zheng (2014), Zheng and Chang (2017) studies, the design is applied on the basis of the comparison of the pretest item's expected total D-optimal value for the examinee in this study. Firstly, when an examinee reaches the seeding position, the D-optimal value for each item is calculated using the ability of the active and past examinee previously took the same pretest item, and then the item having the maximum D-optimal value is selected (Zheng, 2014).

Bayesian-D optimality design. The Bayesian-D optimality design is a restructured version of the D-optimal design. Similar to the method of comparing D-value, the item having the largest Bayesian D-optimality value (as explained Equation 11 in the Model, Concept, and Notations section) is selected by comparing all the pretest items for the examinee. In this method, the item is chosen among the pretest items according to which item's expected contribution will be maximum by the examinee (van der Linden, & Ren, 2015).

Parameter estimation methods. In this part, Joint Maximum Likelihood and One EM Cycle are introduced.

Joint Maximum Likelihood. JML is a technique that simultaneously estimates both ability and item parameters by maximizing likelihood function in IRT, commonly preferred especially in the early stages of IRT. It is almost never used in the online calibration literature except for one study (Verschoor et al., 2019) as a parameter estimation method. It was conducted in a different calibration design from this study. As previously described, JML is a two-stage iterative process. In the first stage, the item parameters are fixed and the ability is estimated, whereas in the second stage, the abilities calculated in the first stage are fixed and the item parameters are estimated (Hambleton, Swaminathan, and Rogers, 1991). As the final abilities of the examinees (estimated using operational items) are known in this study design, in other words these are fixed, the second stage of this method is employed for item parameter estimation. The disadvantages of JML method are stated in the literature; making biased estimations in some conditions (de Gruijter, 1990; Drasgov, 1989; Holland, 1990) and failure to estimate parameters when all answers are true or false (Embretson & Reise, 2000). Although these shortcomings, it is included in this study because it is simple, easy to applicable and programmable, efficient to calculate (Embretson & Reise, 2000), and quite fast compared to MML based methods (Verschoor et al., 2019).

One EM Cycle. OEM is a parameter estimator commonly used in online calibration studies includes one EM cycle. In E step, this approach calculates the expected posterior distribution of abilities from administered operational items. In the next step M, it estimates the test item parameter using this posterior distribution obtained in step E to maximize the Marginal Maximum Likelihood function (Wainer

& Mislevy, 1990). It is also a more suitable option instead of MEM which contains statistically intensive process when the parameter update sample is small.

Online Calibration Computer Program

This computer program can simulate all online calibration procedure as explained in the Simulation Study Design section. It allows the control of IRT models (1-PL and 2-PL), CAT elements (first and next item selection, ability estimation, and termination rule), the fundamental (pretest item selection and pretest item parameter estimation) and administration (seeding location, termination rule, and sample size) elements of online calibration and extra features related to the process (i.e., the number of items to be administrated to each examinee and the number of the update sample). As mentioned above, it was developed using the 'Rcpp' package in the R program. The reason why this program was written by the researcher is the absence of commercial or free software to simulate the online calibration process. The development stages of the program are detailed below.

Firstly, a literature review on online calibration was carried out to determine which software (and programming language) simulate the process. There are various commercial and free software and R packages to run CAT simulations; Software: FireStar (Choi, 2009), CATsim (Weiss & Guyer, 2012), and SimulCAT (Han, 2012) and R Packages: catIRT (Nydick, 20014) and catR (Magis, Raiche ve Barrada, 2018). However, none of them allow the simulation of online calibration. Only SimulCAT allows random administration of the pretest items during the operational CAT test, but it does not estimate any pretest parameter. It is seen that all simulation studies are carried out with the computer programs written by the researchers. They have simulated their studies with Fortran, C and C++ program languages and Matlab and R programs. Since the online calibration process has computationally intensive procedures, the author of this study has written his computer program using the "Rcpp" package, which provides the integration of the open-source and free R program and the C++ programming language and enables the easy and fast implementation of high-performance computation (Eddelbuettel, 2013). The program has been developed to consist of a 3-tier theoretical structure and technical functions that support this structure. The development scheme of the online calibration program is presented in Figure 1.

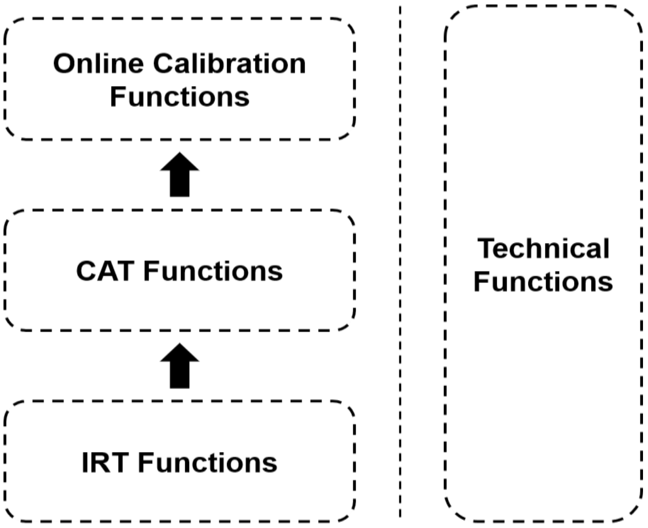


Figure 1. The development scheme of online calibration program

As can be seen in Figure 1, IRT, CAT, and online calibration functions constitute the structure from the lowest level to the highest level, respectively. The example codes of the program for each level are presented in Appendix A. IRT functions are basic and perform the following operations based on theory;

- Calculating the probability of responding to an item,
- Calculating item information (Fisher Information),
- Computing Likelihood function,
- Calculating the ability of examinee and its standard error,
- Generating item response.

CAT functions cover the entire operational test process, from the administration of the first item/s to the termination of the test. These functions perform the following operations.

- Selecting the first and following item/s,
- Stopping the test,
- Simulating the operational Cat test based on the above functions for an examinee.

Online calibration functions simulate the entire calibration process as mentioned in the Literature Review section including operational CAT test and perform the following operations;

- Computing D-optimal Value,
- Selecting the pretest items (random, MFI, D-optimal value design, and Bayesian-D optimality design),
- Estimating the parameter of the pretest items (JML and OEM),
- Simulate the online calibration process based on the above functions for a pretest item bank.

Technical functions are the infrastructure of other functions in the program, such as integrating, adding and transforming the data types used in the “Rcpp” package and performing some calculations (i.e., calculating the integral). The ‘RcppArmadillo’ package (Eddelbuettel, François, Bates, & Ni, 2019), which links the ‘Rcpp’ package with a linear algebra library (Armadillo) in C++, has also been used in the functions (i.e., computing D-optimal value) in which matrix operations are performed. Numerical methods were used in the functions of estimating the ability and pretest item parameters. The Newton-Raphson method and Brent's algorithm (Brent, 1973) were used to determine the maximum likelihood function in the ability estimation functions. The multiple Newton-Raphson method was applied to estimating the pretest item parameters. As aforementioned, the values of the converge threshold and the number of iteration and are determined as 0.001 and 100, respectively for multiple Newton-Raphson iterations.

All simulations were carried out separately on 2 laptops and 6 virtual machines with different features to save time. The laptops have an Intel® Core™ i7 CPU, 2.80GHz, and 16GB of RAM, and Intel® Core™ i5 CPU, 1.70GHz, and 6GB of RAM, respectively. The virtual machines are equipped the same features as threes: Intel Xeon® Platinum 8175 processors (4 virtual cores), up to 3.1GHz, and 16GB of RAM in Amazon Elastic Compute Cloud (EC2; Amazon Web Services, 2019) and Intel® Xeon® E5-2673 processors (2 virtual cores), 2.3GHz, and 8GB of RAM in Azure Virtual Machines (Microsoft Azure, 2019). The R program was run on using Rstudio Desktop and RStudio Server for the laptops and the virtual machines, respectively.

Evaluation Criteria

For each simulation condition, the performance of the tested factors on the accuracy of parameter estimation was evaluated by calculating bias and root mean squared error (RMSE). Bias and RMSE is calculated using the following formulas;

$$Bias_{\eta} = \frac{1}{rep}\frac{1}{m}\sum_{r=1}^{rep}\sum_{i=1}^{m}\left(\hat{\eta}_{r_i}-\eta_{r_i}\right)$$

,

$$RMSE_{\eta} = \sqrt{\frac{1}{rep}\frac{1}{m}\sum_{r=1}^{rep}\sum_{i=1}^{m}\left(\hat{\eta}_{r_i}-\eta_{r_i}\right)^2}$$

where $\hat{\eta}_{r_i} = (\hat{a}_{r_i}, \hat{b}_{r_i})$ and $\eta_{r_i} = (a_{r_i}, b_{r_i})$ are the estimated and true parameters of *i* th item in the *r* th replication. The bias values closer to 0 and the smaller RMSE values indicate higher precision on parameter estimation.

Besides, the cumulative sample size of each pretest item (i.e. N_c) in the online calibration process was recorded and presented in tables and graphs to compare the effectiveness of each item selection method. For this, when the pretest item reached the calibration sample size, the total sample size including examinees from the beginning to the current of the calibration process was recorded for each replication. These sample sizes are sorted from small to large. The mean value of the smallest samples is calculated by dividing the number of replications, then this process is continued until the average of the largest sample number is calculated. If the mean value is decimal, it is rounded to an integer. This data is summarized using graphical representation for each condition.

Chapter 4
Findings and Discussion

The results are presented in the order following the research questions of the study; Comparison of Pretest Item Selection Methods, Parameter Estimation Methods, Sample Size of The Random Calibration Stage, and Calibration Sample Size of Per Pretest Item. Since the simulations process was carried out separately according to 1-PL, and 2-PL model, the results are provided separately for each model under each title.

Comparison Pretest Item selection methods

To compare the effect of pretest item selection methods on parameter precision, bias and RMSE values were calculated for each condition and each parameter. These statistics of each parameter are presented in Table 3 and Table 4 for 1-PL and 2-PL model, respectively. RMSE values of item selection methods grouped according to parameter estimation methods (JML and OEM) for other conditions crossed by the sample size of random calibration phase (n_r = 250, 500, and 1000) and calibration sample size of per pretest item (*N* = 250, 500, and 1000) in Tables 3 and 4 are graphed in Figure 2 for 1-PL model and *b* parameter, Figure 3 for 2-PL model and *a* parameter, and Figure 4 for 2-PL model and *b* parameter.

Table 3

Bias and RMSE of b parameter for Different Item Selection Methods Under 1-PL Model and Different Conditions

n_r	*N*	Item Selection Method	JML		OEM	
			Bias	RMSE	Bias	RMSE
			b	b	b	b
250	250	MFI	-0.0050	0.1710	-0.0049	0.1442
		DOVD	0.0015	**0.1640**	-0.0007	0.1537
		BDOD	-0.0001	0.1725	-0.0010	**0.1424**
	500	MFI	-0.0035	0.1372	-0.0016	0.1033
		DOVD	0.0009	**0.1124**	0.0006	0.1066
		BDOD	0.0040	0.1346	-0.0029	**0.1026**
	1000	MFI	-0.0004	0.1160	0.0013	0.0772
		DOVD	-0.0005	**0.0806**	0.0015	**0.0761**
		BDOD	-0.0002	0.1174	0.0009	0.0785
500	250	MFI	0.0010	0.1623	0.0006	**0.1390**
		DOVD	0.0009	**0.1570**	0.0015	0.1507
		BDOD	0.0007	0.1661	0.0042	0.1399
	500	MFI	-0.0005	0.1353	-0.0027	**0.1009**
		DOVD	0.0005	**0.1120**	-0.0011	0.1070
		BDOD	0.0011	0.1349	0.0031	0.1023
	1000	MFI	0.0029	0.1142	-0.0002	**0.0763**
		DOVD	-0.0012	**0.0796**	0.0028	0.0777
		BDOD	0.0007	0.1167	0.0026	0.0774
1000	250	MFI	0.0057	0.1599	-0.0036	0.1548
		DOVD	0.0047	0.1647	0.0071	0.1562
		BDOD	-0.0066	**0.1584**	0.0065	**0.1540**
	500	MFI	0.0047	0.1267	-0.0019	0.1063
		DOVD	0.0017	**0.1108**	0.0063	0.1112
		BDOD	0.0003	0.1249	0.0048	**0.1046**
	1000	MFI	0.0040	0.1115	-0.0023	**0.0761**
		DOVD	0.0029	**0.0773**	0.0040	0.0777
		BDOD	0.0002	0.1118	0.0032	0.0766

Note. The best values of three item selection method are printed in boldface.
RMSE= Root Mean Squared Error. MFI=Maximum Fisher Information,
DOVD= D-optimal Value Design, BDOD= Bayesian-D optimality design

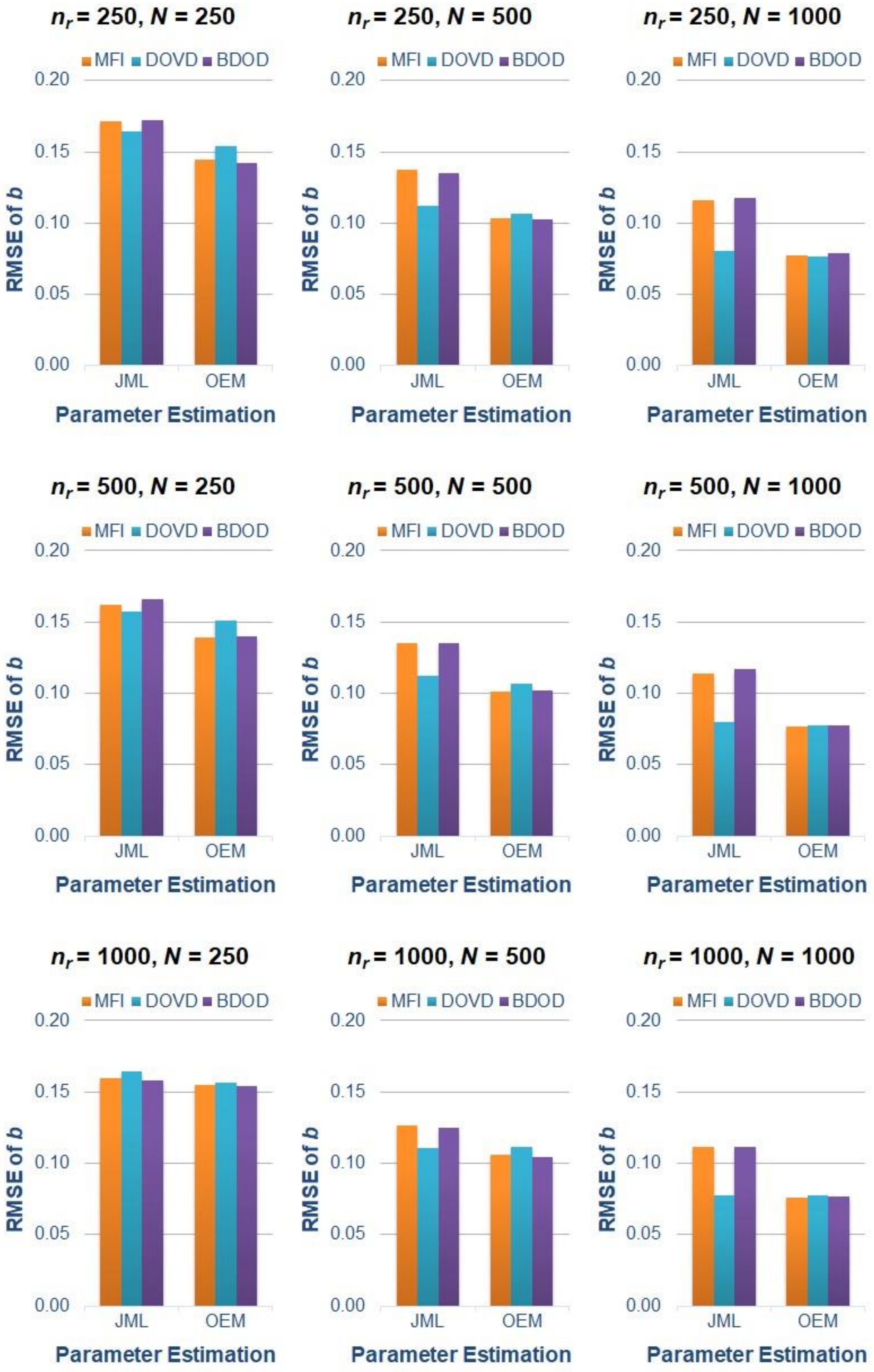


Figure 2. RMSE of *b* parameter for different item selection methods under 1-PL model and different conditions

As can be seen from Table 3, the bias of *b* parameter for MFI, DOVD, and BDOD ranged from -0.0050 to 0.0057, from -0.0012 to 0.0071, and from -0.0066 to 0.0065, respectively. Table 3 and Figure 2 shows that DVOD had the best performance in terms of RMSE under all conditions except one condition (n_r = 1000 and *N* = 250) for JML. MFI and BDOD performed poorly in most conditions, whereas for OEM, MFI, and BDOD worked better than DVOD under most conditions. What is interesting about the result is in Table 3 that MFI and BDOD produced very close values almost all conditions especially for OEM. When OEM is used as the parameter estimation method, the performance of all three item selection methods improved and approached each other as the calibration sample sizes increased. Accordingly, Figure 1 clearly shows that the results of all item selection methods are very similar under all three n_r conditions when the calibration sample size of per pretest item is large (i.e. *N* = 1000) for OEM. With the increase in sample size for JML, a similar trend was observed in terms of performances as in OEM, but the difference across different item selection methods remains.

Table 4

Bias and RMSE of b parameter for Different Item Selection Methods Under 2-PL Model and Different Conditions

n_r	*N*	Item Selection Method	JML				OEM			
			Bias		RMSE		Bias		RMSE	
			a	b	a	b	a	b	a	b
250	250	MFI	-0.0553	0.0104	**0.2074**	**0.2825**	-0.1027	-0.0023	**0.2043**	**0.2531**
		DOVD	-0.0731	-0.0096	0.2091	0.3344	-0.0934	-0.0114	0.2182	0.3009
		BDOD	-0.0817	-0.0055	0.2088	0.3119	-0.0885	-0.0020	0.2075	0.2898
	500	MFI	-0.0697	0.0022	**0.1617**	**0.1922**	-0.1197	0.0014	0.1815	**0.1904**
		DOVD	-0.0765	-0.0021	0.1694	0.2481	-0.1017	0.0000	0.1682	0.2409
		BDOD	-0.0786	0.0059	0.1733	0.2413	-0.0922	0.0012	**0.1628**	0.2104
	1000	MFI	-0.0797	0.0049	0.1396	**0.1616**	-0.1293	-0.0007	0.1695	**0.1469**
		DOVD	-0.0852	0.0014	0.1533	0.2112	-0.0989	0.0026	**0.1393**	0.1752
		BDOD	-0.0852	0.0030	**0.1392**	0.1970	-0.0979	0.0044	0.1394	0.1756
500	250	MFI	-0.0611	0.0012	0.2121	**0.3165**	-0.0990	0.0045	**0.2059**	**0.2645**
		DOVD	-0.0742	-0.0158	0.2243	0.3517	-0.0892	-0.0015	0.2110	0.3016
		BDOD	-0.0714	0.0024	**0.2095**	0.3167	-0.0876	-0.0040	0.2091	0.3030
	500	MFI	-0.0723	-0.0054	0.1706	0.2282	-0.1193	0.0030	0.1881	**0.2115**
		DOVD	-0.0762	-0.0017	0.1785	0.2538	-0.0974	-0.0033	0.1696	0.2248
		BDOD	-0.0814	0.0010	**0.1631**	**0.2272**	-0.0944	0.0065	**0.1682**	0.2169
	1000	MFI	-0.0767	-0.0010	**0.1370**	**0.1561**	-0.1366	-0.0035	0.1782	**0.1561**
		DOVD	-0.0827	-0.0045	0.1516	0.2068	-0.0990	-0.0005	0.1421	0.1728
		BDOD	-0.0807	-0.0024	0.1461	0.1926	-0.0965	0.0017	**0.1376**	0.1621
1000	250	MFI	-0.0689	-0.0026	**0.2062**	0.3304	-0.0891	0.0025	**0.2016**	0.2893
		DOVD	-0.0747	0.0038	0.2112	0.3481	-0.0897	-0.0030	0.2108	**0.2879**
		BDOD	-0.0736	-0.0089	0.2085	**0.3283**	-0.0930	-0.0040	0.2105	0.2956
	500	MFI	-0.0710	-0.0056	0.1701	**0.2247**	-0.1092	-0.0001	0.1777	**0.1867**
		DOVD	-0.0804	-0.0025	**0.1691**	0.2631	-0.0944	0.0041	0.1642	0.2131
		BDOD	-0.0813	0.0000	0.1693	0.2650	-0.0982	-0.0036	**0.1636**	0.2133
	1000	MFI	-0.0794	-0.0031	0.1471	**0.1800**	-0.1360	0.0016	0.1760	**0.1433**
		DOVD	-0.0805	-0.0050	**0.1389**	0.2035	-0.0990	-0.0017	0.1417	0.1812
		BDOD	-0.0831	0.0008	0.1446	0.2125	-0.0958	-0.0022	**0.1382**	0.1722

Note. The best values of three item selection method are printed in boldface. RMSE= Root Mean Squared Error. MFI=Maximum Fisher Information, DOVD= D-optimal Value Design, BDOD= Bayesian-D optimality design

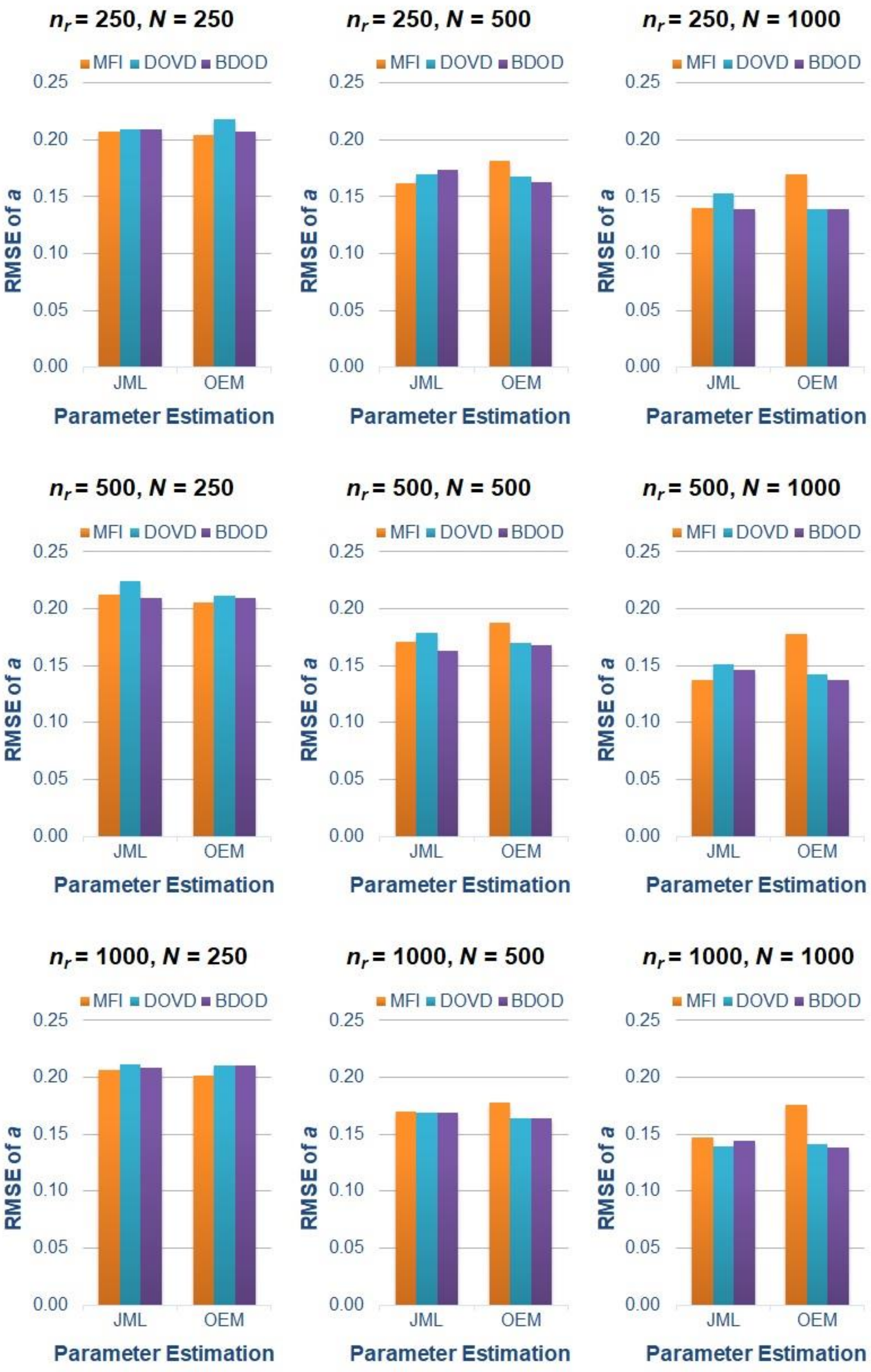


Figure 3. RMSE of *a* parameter for different item selection methods under 2-PL model and different conditions

From Table 4, the bias of *a* parameter for MFI, DOVD, and BDOD ranged from -0.1366 to -0.0553, from -0.1017 to -0.0731, and from -0.0982 to -0.0714, respectively. In terms of the accuracy of *a* parameter, the performance of the item selection methods is comparable according to other conditions. There are no outstanding methods for JML in terms of the lowest RMSE value under the nine conditions. Performance ranking of the methods for JML is MFI (4 out of 9), BDOD (3 out of 9) and DOVD (2 out of 9). However, since DVOD has the highest RMSE values, it is less efficient than other methods. The performance of BDOD is remarkable when OEM was used as the parameter estimation method and especially *N* = 500. Looking at Figure 3 for the OEM, it is apparent that DVOD consistently worked worse than the others when calibration sample size of per pretest item is small level (i.e. *N* = 250), and MFI obviously performance poorly when it is medium level (i.e. *N* = 500), and especially large level (i.e. *N* = 1000). The increase in the sample size of the random calibration phase did not cause a clear pattern in the performance of the methods. On the other hand, the methods had lower RMSE values with the increase of calibration sample size of per pretest item, so their performance improved.

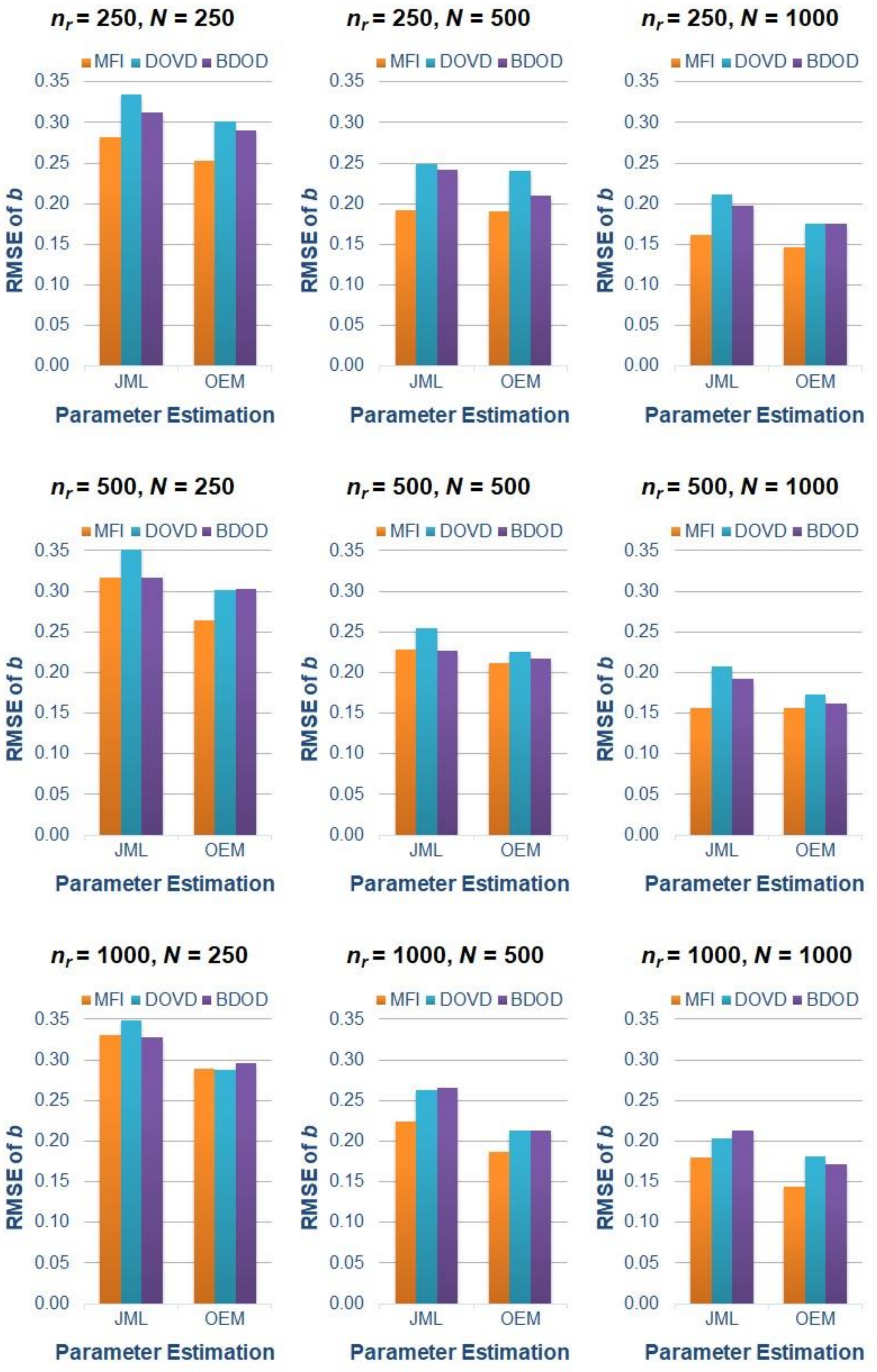


Figure 4. RMSE of *b* parameter for different item selection methods under 2-PL model and different conditions

By browsing Table 4, the bias of *b* parameter for MFI, DOVD, and BDOD ranged from -0.0056 to 0.0104, from -0.0158 to 0.0041, and from -0.0089 to 0.0065, respectively. In terms of RMSE, a point worth noting is that MFI consistently resulted in the best item selection method for both parameter estimation methods as can be seen Figure 4. Besides, DOVD is the least effective method - especially small random calibration sample size (n_r = 250) - since it has the largest RMSE value for both parameter methods. For the increase in both sample size of random calibration phase and calibration sample size of per pretest item, similar results were obtained as the result of *a* parameter.

In summary, it can be said that DVOD is the best pretest item selection method for JML, whereas MFI (as expected) and BDOD are better and preferable for OEM in the 1-PL model. These finding for OEM and the 1-PL model is consistent with Zheng (2014)'s finding that MFI outperforms DVOD for *b* parameter. The reason why MFI and BDOD are close to each other for 1-PL can be explained as the convergence of results due to the use of both information functions and the fact that the ability was estimated only based on *b* parameter. For the 2-PL model, DVOD is the worst in terms of the accuracy of the discrimination parameter. The results of this study do not match the findings of Zheng (2014) that MFI (the examinee-centered method) is the worst outcome among the five item selection methods and Zheng and Chang (2017). Finally, as expected and the 1-PL model, MFI is the best choice among other pretest item selection designs. This can be explained by the fact that MFI has a greater value when the difficulty of the selected item approaches the ability level, as Zheng (2014) states. BDOD performed close to the best performing methods in both 1-PL and 2-PL models. While this method performed better in extreme difficulty parameter than the other methods (MFI, DVOD, and OIRPI) in Zheng and Chang (2017) study and head-to-head with the random method in Chen (2017) study, it lagged behind the Excellence Degree (ED) criterion in He et al. (2020) study and also showed poor performance for the lower discrimination parameters in Zheng and Chang (2017) study.

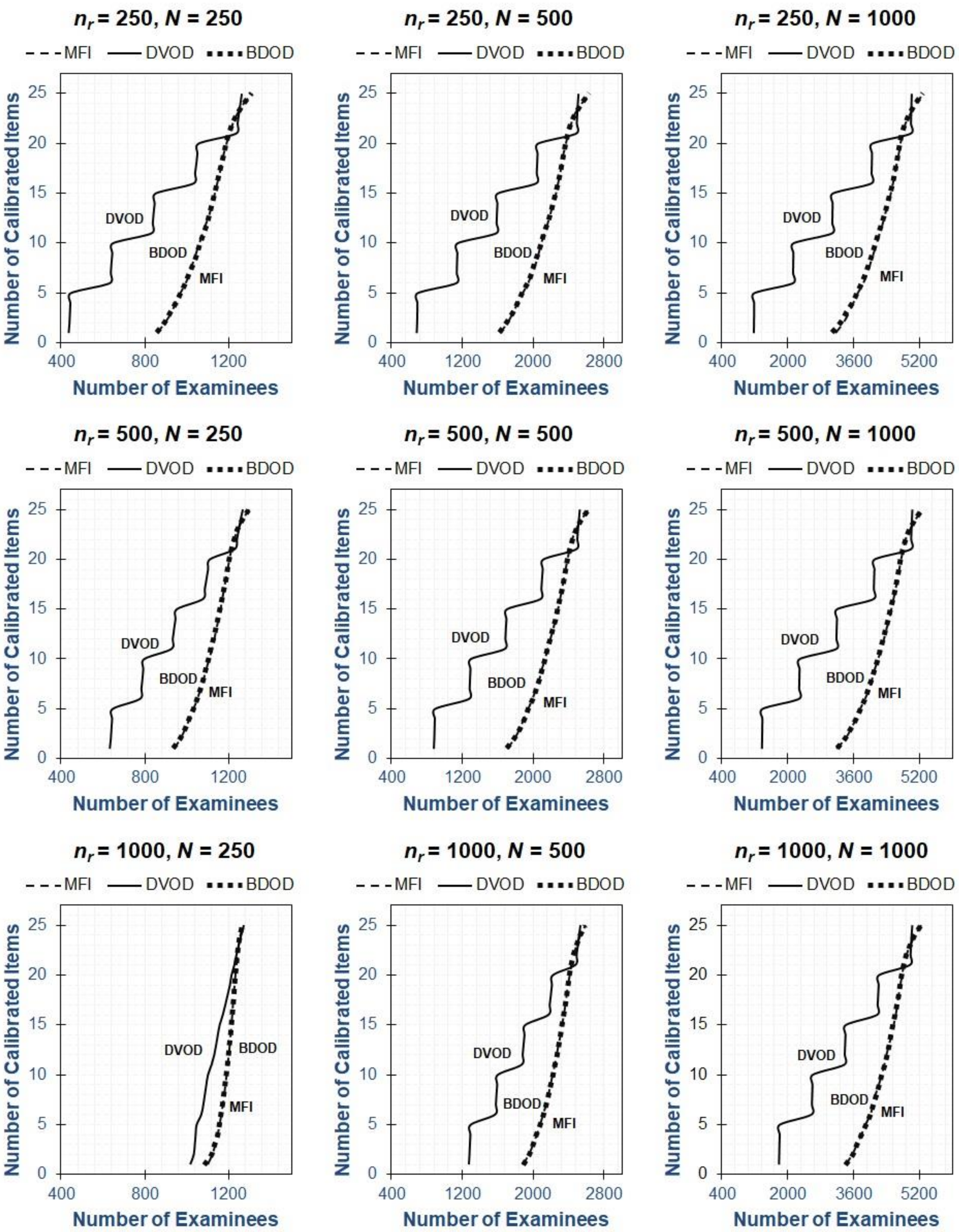
In this study, the cumulative sample size of pretest items was used as a measure of comparison of the effectiveness of item selection methods. For this purpose, the average of simulated examinees required from the first item (starting to retirement of pretest items) to the last item (in other words, for all items or retirement of all pretest items) whose calibration was completed, in other word which exported from pretest item bank, was put on record for each condition. These values for the first and last items in the 1-PL model and 2-PL model are shown in Tables 5 and 6, respectively. In addition, the cumulative sample size for all items is plotted in Figures 5 and 6 for OEM in the 1-PL and the 2-PL model, respectively. Since the results of the pretest item selection methods are very close and showed similar for JML and OEM, the plots of the cumulative sample size for JML is presented in Appendix B and Appendix C for 1-PL and the 2-PL model, respectively.

Table 5

The Cumulative Sample Size of First and Last Item for Different Item Selection Methods Under 1-PL Model and Different Conditions

n_r	N	Item Selection Method	JML		OEM	
			First Item	Last Item	First Item	Last Item
250	250	MFI	856	1313	857	1313
		DOVD	439	1265	438	1264
		BDOD	854	1314	855	1311
	500	MFI	1625	2644	1628	2643
		DOVD	689	2513	688	2514
		BDOD	1630	2645	1618	2638
	1000	MFI	3158	5296	3160	5286
		DOVD	1189	5013	1188	5014
		BDOD	3166	5297	3086	5283
500	250	MFI	942	1297	938	1298
		DOVD	634	1270	634	1270
		BDOD	942	1299	937	1297
	500	MFI	1717	2622	1703	2627
		DOVD	884	2520	884	2520
		BDOD	1700	2624	1702	2621
	1000	MFI	3257	5281	3192	5264
		DOVD	1384	5019	1384	5020
		BDOD	3251	5284	3213	5266
1000	250	MFI	1083	1267	1099	1268
		DOVD	1027	1274	1017	1274
		BDOD	1097	1269	1083	1267
	500	MFI	1891	2593	1885	2591
		DOVD	1278	2529	1277	2530
		BDOD	1891	2596	1887	2591
	1000	MFI	3434	5249	3407	5240
		DOVD	1778	5029	1777	5030
		BDOD	3409	5255	3396	5240

MFI=Maximum Fisher Information, DOVD= D-optimal Value Design,
BDOD= Bayesian-D optimality design



Note. The lines of MFI and BDOD overlap.

Figure 5. The cumulative sample size of pretest items for OEM methods under 1-PL model and different conditions

Table 5 and Figure 5 indicate that DVOD is obviously the first method to initiate the retirement of pretest items with a minimum number of simulated examinees for all pretest estimation methods, the sample size of the random calibration phase and calibration sample size of per pretest item. Moreover, it completed the calibration process/retirement of all items in the pretest item bank at the earliest. On the other hand, MFI and BDOD require more simulated examinees on average in terms of both starting the retirement and completion of the calibration process. In addition to that, their results are very close to each other in terms of cumulative sample size as well as RMSE under all conditions. Even, they have the same average values in some conditions, for example, when n_r = 1000 and N = 500 and 1000 for OEM.

Table 6

The Cumulative Sample Size of First and Last Item for Different Item Selection Methods Under 2-PL Model and Different Conditions

n_r	N	Item Selection Method	JML		OEM	
			First Item	Last Item	First Item	Last Item
250	250	MFI	501	1328	507	1327
		DOVD	440	1271	439	1276
		BDOD	446	1310	446	1313
	500	MFI	813	2678	820	2675
		DOVD	690	2521	690	2540
		BDOD	696	2596	696	2593
	1000	MFI	1423	5381	1426	5353
		DOVD	1190	5022	1189	5030
		BDOD	1196	5109	1196	5118
500	250	MFI	687	1309	689	1309
		DOVD	635	1278	637	1277
		BDOD	646	1305	643	1307
	500	MFI	1004	2650	1010	2666
		DOVD	886	2531	886	2532
		BDOD	896	2599	895	2613
	1000	MFI	1630	5353	1627	5366
		DOVD	1386	5029	1386	5040
		BDOD	1397	5151	1396	5156
1000	250	MFI	1050	1275	1050	1276
		DOVD	1030	1274	1031	1275
		BDOD	1039	1274	1038	1277
	500	MFI	1368	2615	1374	2625
		DOVD	1281	2537	1281	2536
		BDOD	1297	2593	1298	2604
	1000	MFI	1992	5306	2005	5330
		DOVD	1781	5040	1781	5040
		BDOD	1798	5172	1798	5188

MFI=Maximum Fisher Information, DOVD= D-optimal Value Design,
BDOD= Bayesian-D optimality design

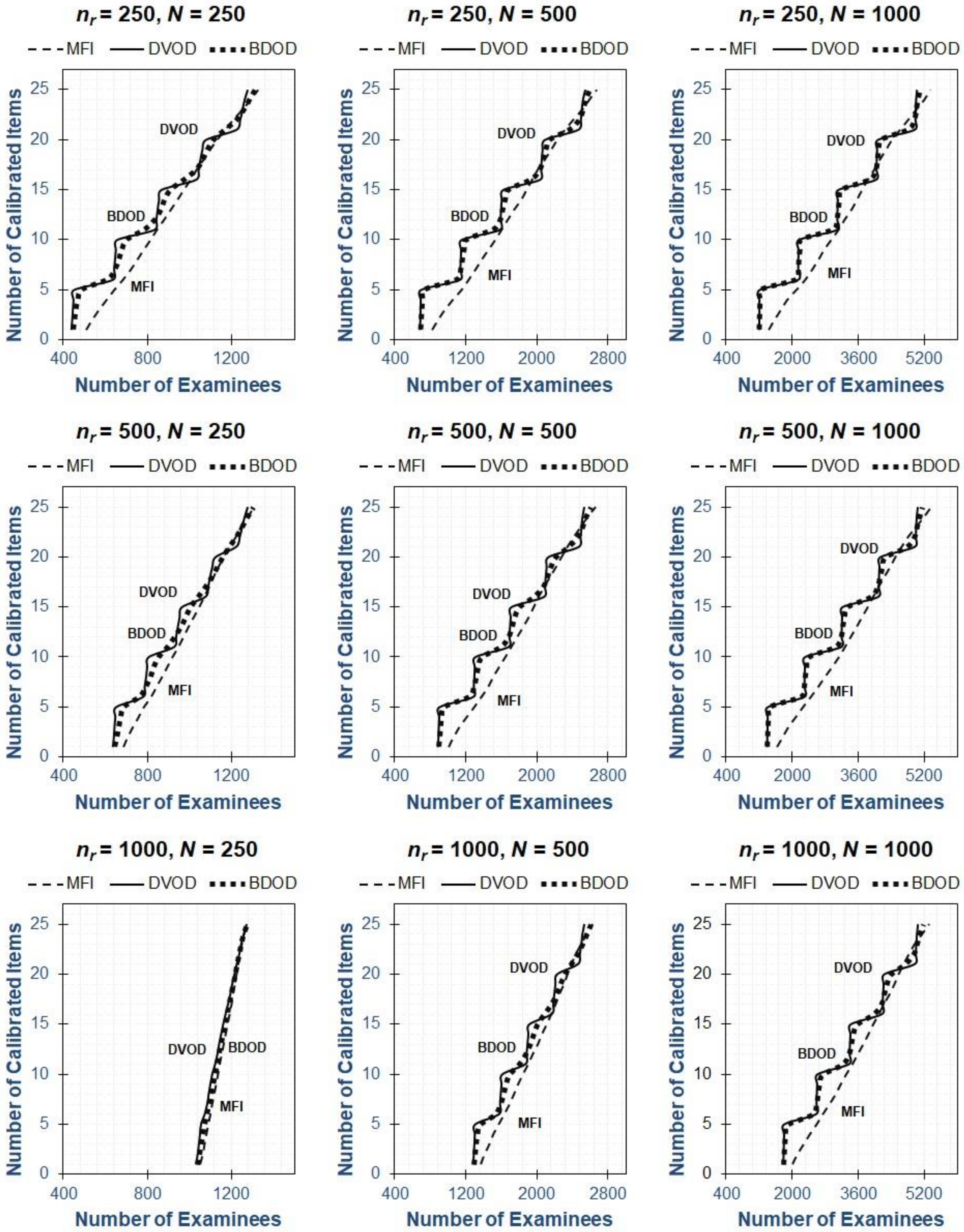


Figure 6. The cumulative sample size of pretest items for OEM methods under 2-PL model and different conditions

As can be seen from the average simulated examinees for the first items in Table 6, DVOD is the first method to retire the items under all conditions at the earliest, as in 1-PL. In terms of this criterion, it can appear from Figure 6 that the results of BDOD are close to the DVOD. MFI is worst because it needs more simulated examinees. In terms of the termination of all pretest items retirement, the three methods can be rank ordered as (from the earliest to the latest): DVOD, BDOD, and MFI. When n_r = 1000 and N = 250, all three pretest item selection methods behaved similarly for both pretest estimation methods.

When comparing pretest item selection methods for the 1-PL and 2-PL model, an interesting finding is that MFI and BDOD methods in the 1-PL model require more simulated examinees than 2-PL model to complete the calibration of the first pretest item. For DVOD, the average number is quite close in both models. To complete the calibration process, all three methods require fewer examinees for 1-PL model than 2-PL model. Another interesting result is that MFI and BDOD need less average simulated examinees with the increasing of random stage sample size and vice versa for DVOD in the conditions when the calibration sample size is the same for 1-PL model. Moreover, MFI showed the same tendency for the 2-PL model. Apart from all these, by comparing the cumulative sample size of the parameter estimation methods with Tables 5 and 6 for the 1-PL model and 2-PL model, it can easily notice that these numbers are close to each other or even the same for JML and OEM.

Summarizing the results in terms of cumulative sample size, DVOD tends to retire early for both the 1-PL model and the 2-PL model. DVOD followed a characteristic curve similar to the progressive stairs that retired five items with the highest D-optimal value each time and then moved on to the next five items. The disadvantage of this is that the pretest items with the highest D-optimal value tend to be selected continuously, regardless of the ability of the next examinee, and other pretest items cannot be selected (Zheng, 2014). As a result, however, other pretest items are administrated after the favourite one’s retirement. For the 2-PL model, BDOD followed the similar characteristic curve as in the study of van der Linden and Ren (2015). Accordingly, it has been observed by Chen (2017) that some pretest items tend to be preferred and selected more than others by BDOD similar to DVOD. MFI is the worst method for this criterion because it is designed as an examinee-centered method rather than an item- centered method that optimizes the calibration of the pretest items. Apart from these, the reason why there is no difference between parameter estimation methods in terms of the cumulative sample size is that this variable is directly related to the item selection method.

Comparison Parameter Estimation Methods

To compare the effect of different parameter estimation (JML and OEM) methods on precision in parameter estimation, average bias and RMSE values were used. In the item selection method comparison section, since bias and RMSE values for both 1-PL (difficulty parameter) and 2-PL model (discrimination and difficulty parameter) are presented in Tables 3 and 4 which include parameter estimation results, no tables are included in this section. Figures 7 through 9 portrays the comparison of parameter estimation methods grouped according to item selection methods (MFI, DVOD, and BDOD) for other conditions crossed by the sample size of random calibration phase (nr = 250, 500, and 1000) and calibration sample size of per pretest item (N = 250, 500, and 1000) in Tables 3 and 4 for *b* parameter in 1-PL model, *a* parameter in 2-PL model, and *b* parameter in 2-PL model.

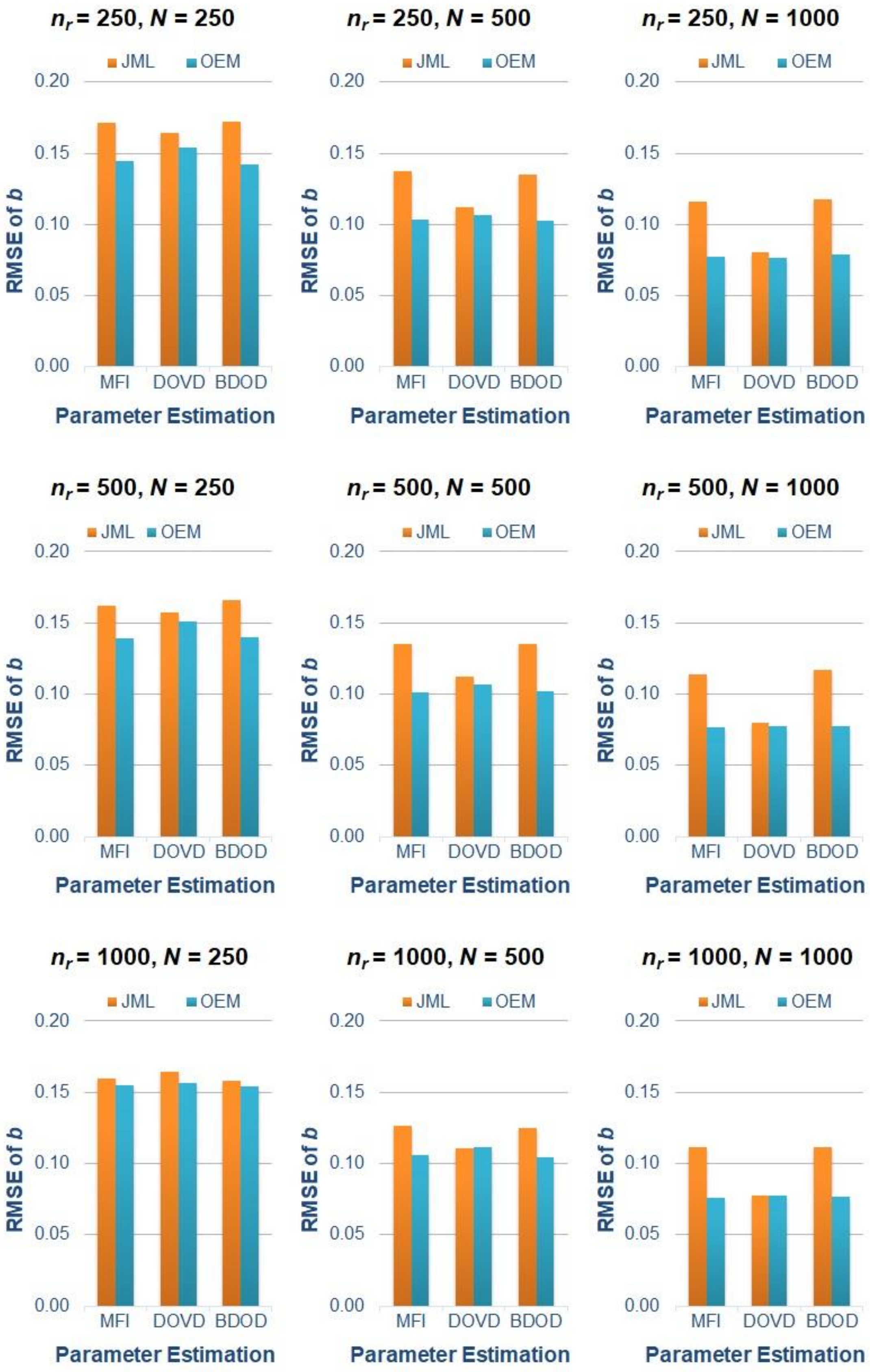


Figure 7. RMSE of *b* parameter for different parameter estimation methods under 1-PL model and different conditions

According to Table 3, the bias of *b* parameter for JML and OEM ranged from -0.0066 to 0,0057 and from -0.0049 to 0.0071, respectively. These two calibration methods tended to overestimate difficulty parameters, as stated by the positive biases in Table 3. When the RMSE values as a measure of parameter accuracy are compared, it is seen from Table 3 and Figure 7 that the OEM method is a clear best method for *b* parameter. The differences of JML and OEM are more obvious in the conditions of MFI and BDOD for medium (*N* = 500) and large (*N* = 1000) calibration sample size. When DVOD used as the item selection method, although the difference between JML and OEM is small, JML performance is generally behind the OEM. Only under two conditions (when n_r = 1000 and *N* = 500 and 1000 for DVOD), JML produced slightly lower RMSE values than OEM but the difference is very small.

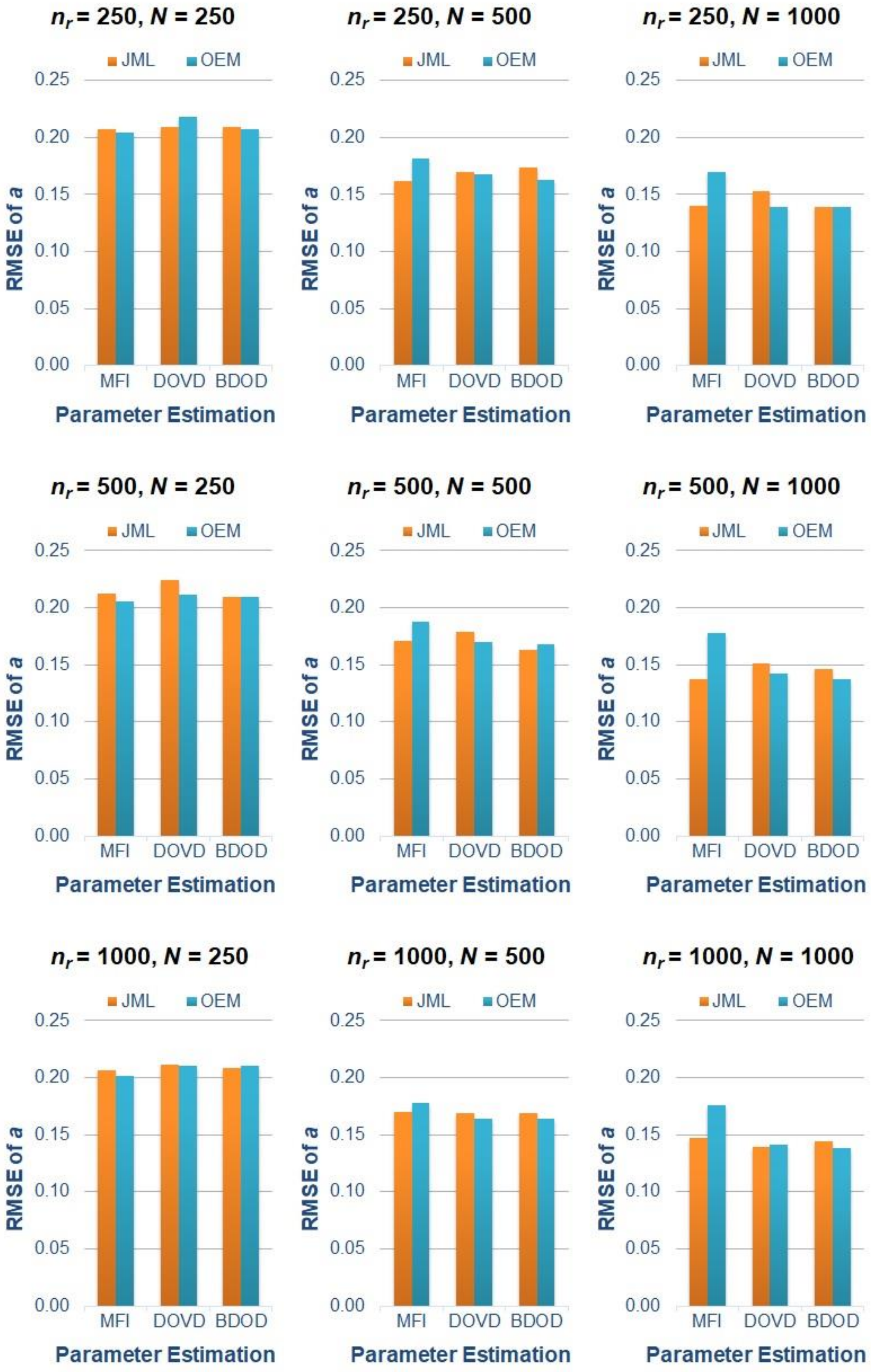


Figure 8. RMSE of *a* parameter for different parameter estimation methods under 2-PL model and different conditions

In Table 4, the bias of *a* parameter for JML and OEM ranged from -0.0852 to -0,0553 and from -0.1366 to -0.0876, respectively. Between these two methods, JML yields smaller bias than OEM. Besides, as can be seen from the negative biases in Table 4, these methods underestimated the discrimination parameter under all simulated conditions. In terms of RMSE, these two methods provided comparable results according to pretest item selection methods. For MFI, OEM was found slightly more efficient than JML, when the calibration sample was small (N = 250). In contrast, JML is more sensitive than OEM as indicated by lower RMSE values when the calibration sample was medium (N = 500) and large (N = 1000). For DVOD, JML generated less RMSE only two conditions - when both random phase and calibration sample size is small (n_r = 250 and N = 250) and large (n_r = 1000 and N = 1000). In other conditions, OEM performed better than JML. Lastly, for BDOD, OEM shows better performance than JML. Besides, JML performed better in a very interesting pattern. If one looks at Figure 8 diagonally from top right to bottom left, it can be seen that JML behaved better than OEM. Summarized in terms of recovering discrimination parameter, when MFI is set aside, it can be seen that the two parameter estimation methods generally produce close values regardless of whether they perform better or worse.

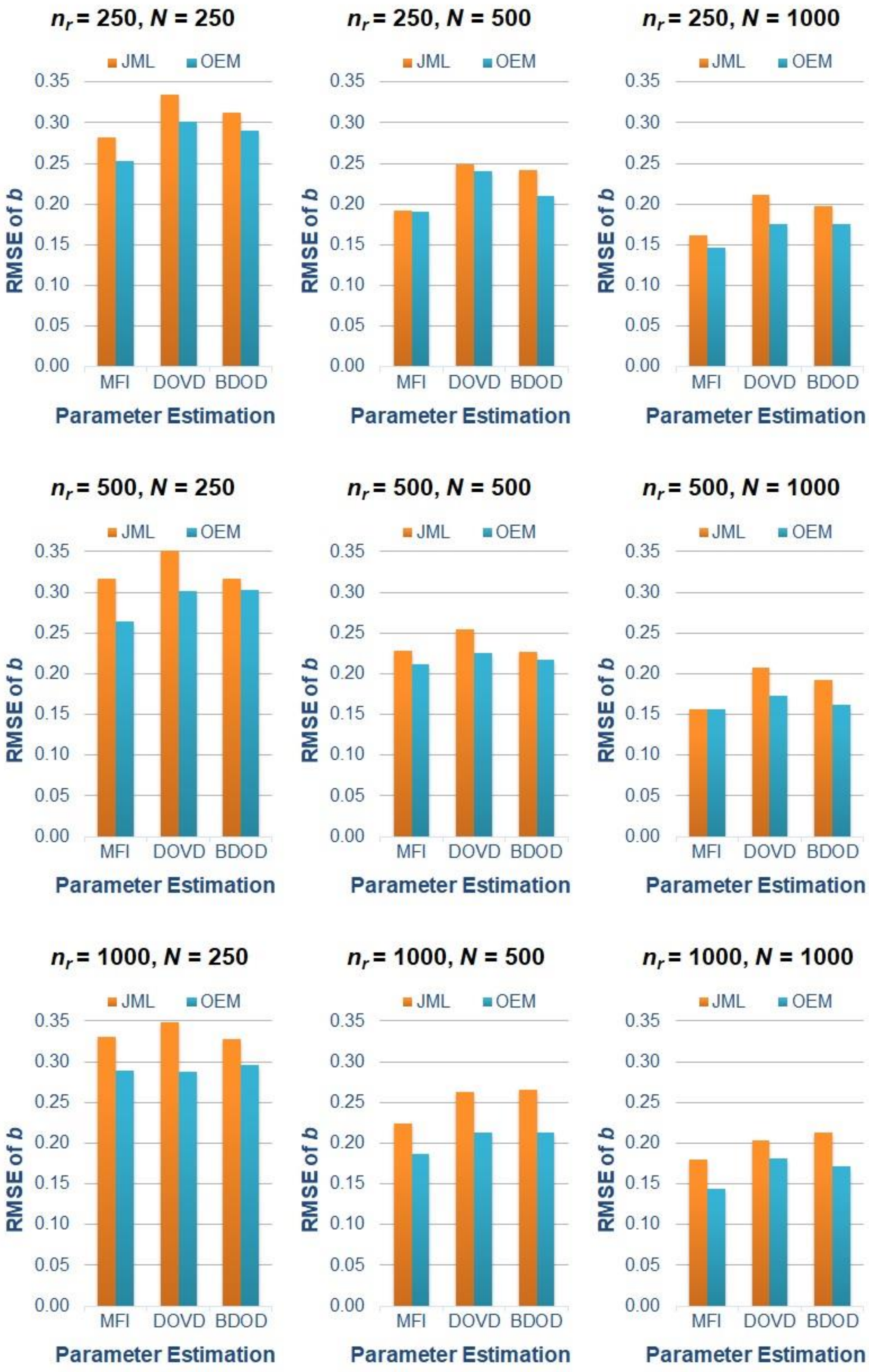


Figure 9. RMSE of *b* parameter for different parameter estimation methods under 2-PL model and different conditions

From Table 4, the bias of b parameter for JML and OEM ranged from -0.0158 to 0,0104 and from -0.0114 to 0.0065, respectively. Accordingly, the bias results of the OEM are closer to 0. Even, the bias value is 0 when n_r = 500 and N = 500. Both JML and OEM tended to underestimate difficulty parameters, as stated by the negative biases in Table 4. Compared to the performance of parameter estimation methods in terms of RMSE with Table 4 and Figure 9, OEM produced smaller values in almost all conditions and the differences between the methods is more obvious. In other words, JML consistently showed worse performance than OEM. When n_r = 500 and N = 1000, two parameter estimation methods exhibited the same performance.

To summarize the results in terms of parameter estimation methods, OEM is yielded better difficulty parameter recovery for both the 1-PL model and the 2-PL model. For the recovering of the discrimination parameter, OEM performed slightly better than JML. OEM is based on MMLE and has been used in studies in the online calibration literature (Ban et al., 2001; He et al., 2017; He et al., 2020; Zheng, 2014; Zheng & Chang, 2017). These studies have also demonstrated the effectiveness of the pretest calibration method in terms of parameter accuracy and speed. Zheng and Chang (2017) proposed OEM for the 1-PL model because it has similar parameter recovery performance as MEM (which is the best method) with shorter processing time. They also proposed the use of the OEM method when the pretest item selection method is not examinee-centered and pretest items with larger discrimination parameters if the processing time is important. Likewise, Zheng (2014) stated that it is worse when MFI is the pretest item selection method for the precision of the discrimination parameter. However, in this study, OEM performed well in examinee-centered methods (MFI). He et al., (2017) explained that the reason why the OEM method works well is that it is not affected by the measurement errors within theta estimates by integrating the latent abilities out in the M step.

Verschoor et al. (2019) used JML as a parameter estimation method in online calibration (they called On-the-Fly Calibration). However, it was used to calibrate the item pool instead of the individual estimation of each item parameter. With this aspect, JML was used for the first time as a parameter estimation method in online calibration. It is known that JML has some disadvantages when used as parameter estimation method in fixed length tests; biased parameter estimation, have questionable standard errors, and do not improve the accuracy of calibration while increasing the number of test takers is expected to increase (Embretson & Reise, 2000; Holland, 1990). However, in this study, instead of two stages of JML, a single stage was employed and, in this stage, the ability parameters obtained at the end of the CAT test were fixed. As a result, the biases that may arise from ability estimation have been slightly reduced. Although it maintains its genetic disadvantages, it performed a bit better or closer to OEM in the estimation of the difficulty parameter for the 1-PL model (when the pretest item selection method is DVOD) and the discrimination parameter for the 2-PL model (when the pretest item selection method is BDOD).

When the same conditions are compared in Table 4, the values of *a* parameter are greater than the values of *b* parameter in terms of bias. Conversely, for RMSE, the values of *b* parameter are greater than the values of *a* parameter. This similar results in terms of RMSE have been seen in the studies of Ban et al. (2001), He et al. (2017), and Wang and Xu (2015). He et al. (2017) explained the reason for this by the fact that the discrimination parameter distributed a narrower range (from 0.45 to 2.26) while the difficulty parameter distributed a wider range (from -3.17 to 3.11). Besides, biases of *b* parameters under all conditions are very close to 0. Regardless of the 2-PL model, this finding is consistent with the results of the 1-PL model. It can be also seen from the data about RMSE of *b* parameters in Table 3 and Table 4 that the values are lower in the 1-PL model than the 2-PL model under the same conditions. The possible reason for this is that both the a and the *b* parameters need to be calibrated for the 2-PL model, while only the *b* parameter is calibrated for the 1-PL model. Similarly, the difference between JML and OEM was found clearer in the 2-PL model than the 1-PL model.

To compare the effect of different parameter estimation (JML and OEM) methods on precision in parameter estimation at the parameter level, the average biases of the parameters were calculated for each condition and each parameter level. For this purpose, the parameters of all calibrated 2500 pretest items (25 pretest items were calibrated in each replication; *m x rep* = 25 x 100 = 2500) are used. The bias values against the parameter levels were illustrated in Figure 10 for the 1-PL model and *b* parameter, Figure 11 for the 2-PL model and *a* parameter, and Figure 12 for the 2-PL model and *b* parameter.

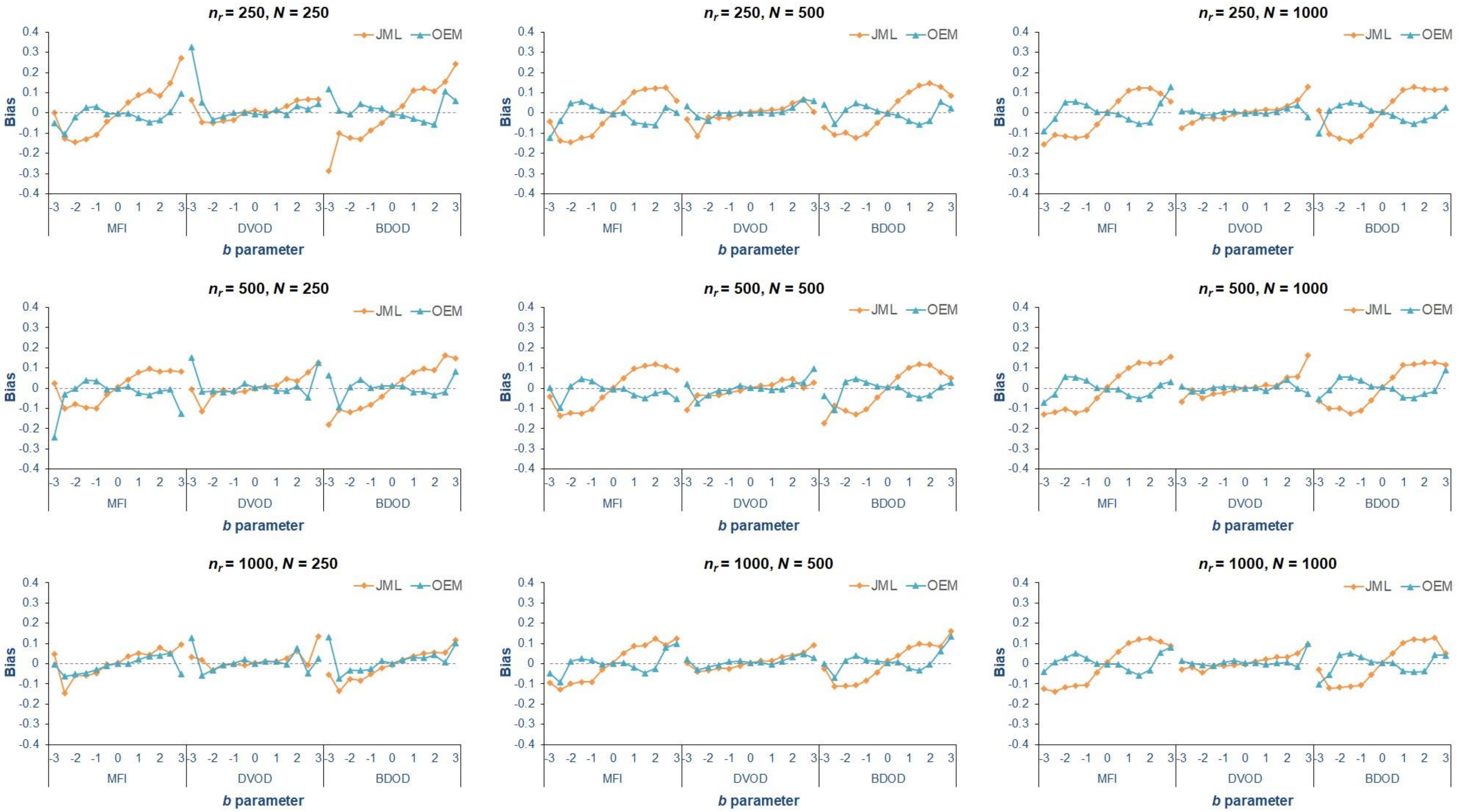


Figure 10. Bias of *b* parameter at each difficulty level for different parameter estimation methods under 1-PL model and different conditions

The effect of the parameter estimation methods on parameter accuracy at each difficulty level in terms of bias for the 1-PL model and *b* parameter is comparable for the pretest item selection methods. According to Figure 10, for MFI and BDOD, JML and OEM performed opposite fluctuation to each other at different difficulty levels except in one condition (n_r = 1000 and N = 250, see the left-bottom in Figure 11). In general, JML underestimates easy items while overestimates hard items. The biases reached a negative and positive peak at around b=-2 and around b=+2, respectively. Contrary to expectations, JML outperformed at negative outliers. However, it performed irregularly in positive outliers. OEM tended to underestimate both very easy and hard pretest items and overestimate both easy and very hard pretest items. For DVOD, JML and OEM performed similar almost all conditions; underestimate for easy pretest items and vice versa for difficult pretest items. They outperformed as indicated by biases close to 0.

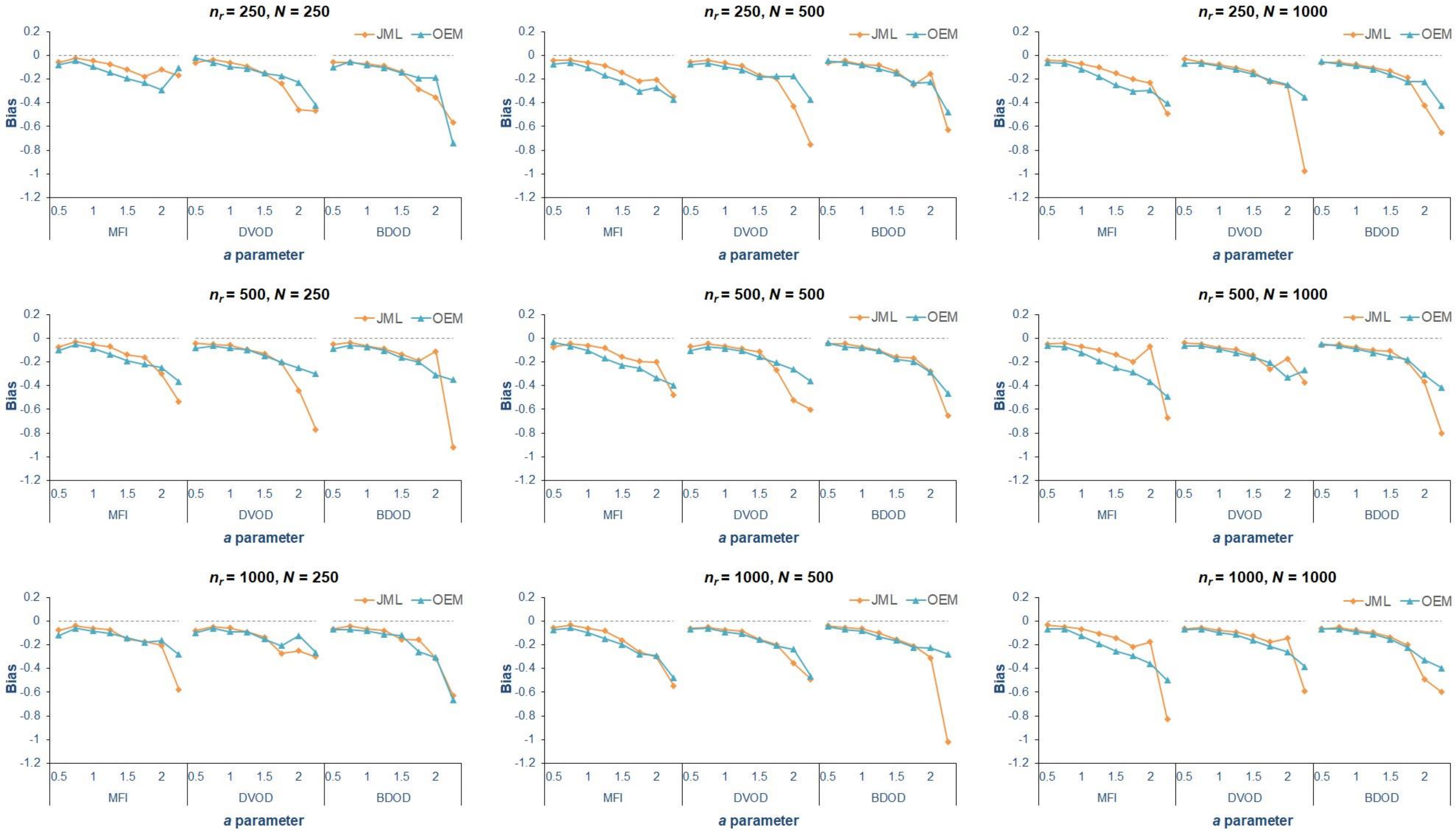


Figure 11. Bias of *a* parameter at each discrimination level for different parameter estimation methods under 2-PL model and different conditions

From Figure 11, the performance of the parameter estimation methods for the 2-PL model and *a* parameter are very similar in almost all conditions. As aforementioned and seen in Table 4, the discrimination parameters are underestimated by these methods. When the discrimination level increased from low to high, (negative) bias values produced by both parameter calibration methods increase moderately. This means that they are less sensitive for pretest items with higher discrimination value. Moreover, JML deviated more than OEM for high discriminating items. However, for larger *a* parameter values, RMSE value produced by OEM increased with an increasing sample size when MFI is used as the pretest item selection method. Zheng and Chang (2017) also reported similar behaviour in their study. Apart from that, when the MFI is used as the pretest item selection method, it is seen again that JML provides more accurate parameter estimation than OEM.

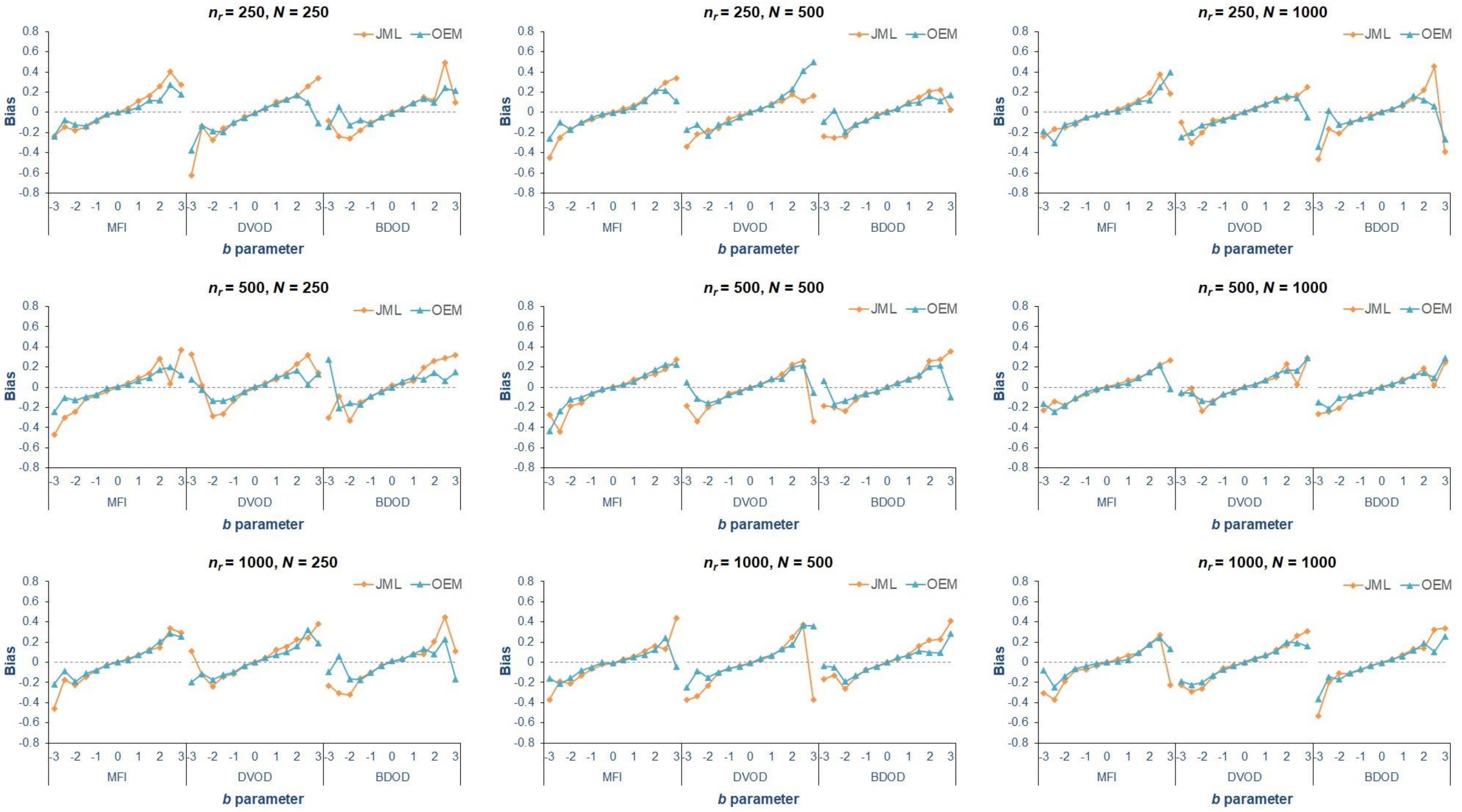


Figure 12. Bias of *b* parameter at each difficulty level for different parameter estimation methods under 2-PL model and different conditions

Figure 12 indicates that the results of the parameter estimation methods on the accuracy of *b* parameter for the 2-PL model are consistently similar under all simulated conditions. When very easy and very difficult items (from at *b* =-3.0 to at *b* = -2.5 and from at *b* = 2 to at *b* = 3) put aside, both calibration methods underestimate easy items while they overestimate hard items, such as JML's performance in the 1-PL model. Medium difficulty items (at around b=0) yielded better parameter accuracy as in the 1-PL model. For very easy and very hard pretest items, JML and OEM could not maintain their characteristic trend, but JML's performance was more deviated.

Comparison Sample Size of The Random Calibration Stage

To compare the effect of the sample size of the random calibration stage (250, 500, 1000) on precision in parameter estimation, average RMSE values were used. This effect was investigated separately for item selection methods and parameter estimation methods, respectively. As explained above, no tables are included in this section. RMSE values of pretest item selection methods at these sample size for other conditions crossed by the parameter estimation methods and calibration sample size of per pretest item (*N* = 250, 500, and 1000) in Tables 3 and 4 are plotted in Figure 13 for the 1-PL model and *b* parameter, Figure 14 for the 2-PL model and *a* parameter, and Figure 15 for the 2-PL model and *b* parameter.

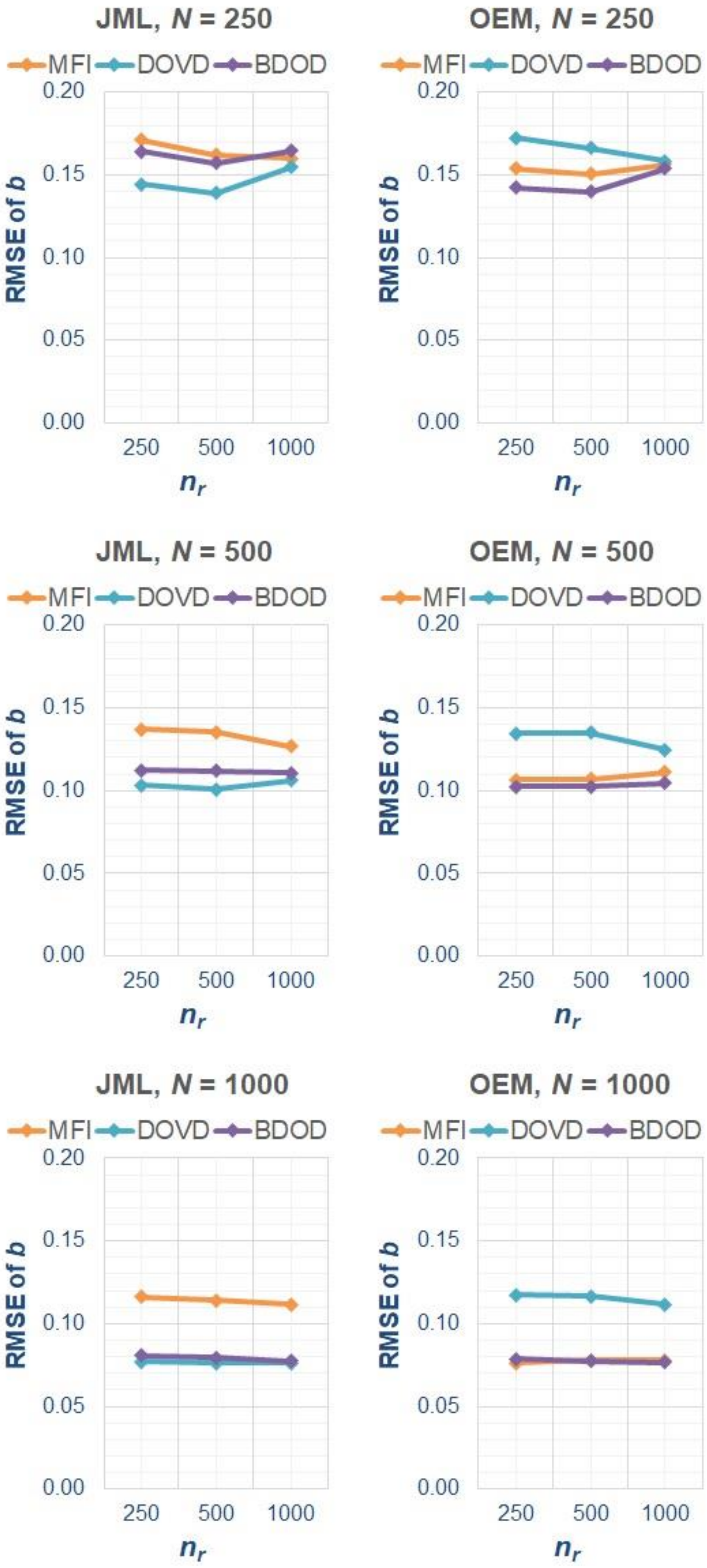


Figure 13. RMSE (*b* parameter) of the item selection method for sample size of the random calibration stage under 1-PL model and different conditions

The effect of the random phase sample size via the pretest item selection methods on parameter accuracy for the 1-PL model and *b* parameter differs according to other simulated conditions, as can be seen from Table 3 and Figure 13. However, their performance is very close to each other for different sample sizes of the random stage under most conditions. Only, the change between sample sizes of the random stage more noticeable when the calibration sample is small (*N* = 250). MFI showed a characteristic pattern when the parameter estimation method is JML. For these conditions, RMSE value decreased slightly as the sample size of the random phase increased from 250 to 1000. Similar trends are seen for DVOD and BDOD when *N* = 1000 and for DVOD when the parameter estimation method is OEM and *N* = 1000. Apart from these conditions, irregular trends are seen when the random phase sample size increases. For example, it is shown a decrease from 250 to 500, followed by an increase from 500 to 1000 for some conditions (for DVOD when JML used as calibration method and *N* = 250 and 500, for BDOD when JML and OEM used as calibration method and *N* = 250). Moreover, as n_r = 1000, the performance of the methods in the small (*N* = 250) and medium (*N* = 500) calibration sample sizes approached each other.

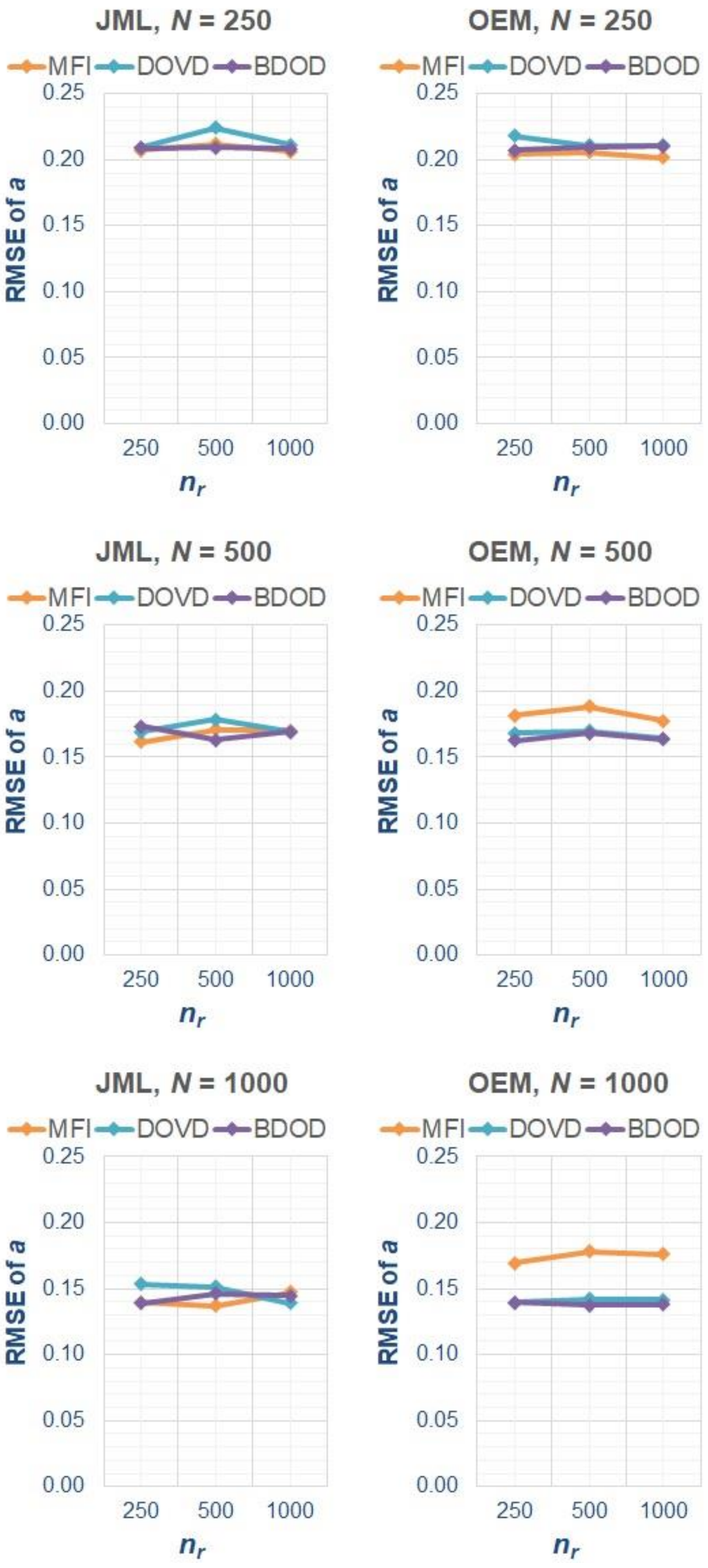


Figure 14. RMSE (*a* parameter) of the item selection method for sample size of the random calibration stage under 2-PL model and different conditions

By browsing Table 4 and Figure 14, with the increasing of random phase sample size under the parameter estimation methods, the performance of the pretest item selection methods for 2-PL model and *a* parameter has aberrant trend as follows in most conditions; an increase from small (n_r = 250) to medium (n_r = 500) followed by a fall from medium (n_r = 500) to large (n_r = 1000). However, these increases and decreases are very small as in the 1-PL model. The conditions in which calibration accuracy of discrimination parameter improves (RMSE decreases) as the sample size of random stage increases when calibration sample sizes are small (*N* = 250) and medium (*N* = 500) for OEM and JML, respectively. In addition, the performance of the methods converges under all calibration sample sizes for the large sample size of the random stage (n_r = 1000) and JML.

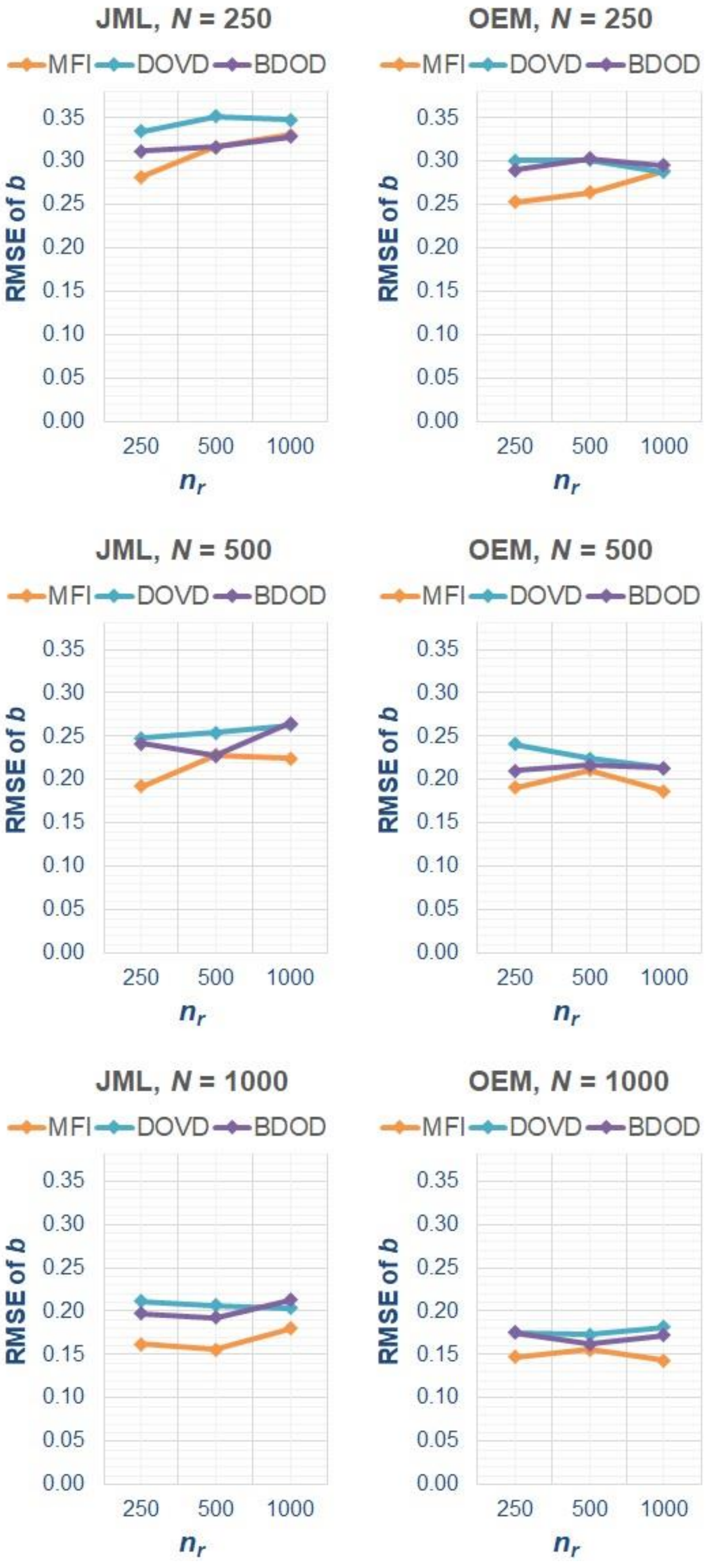


Figure 15. RMSE of (*b* parameter) of the item selection method for sample size of the random calibration stage under 2-PL model and different conditions

As Table 4 and Figure 15 show, depending on the increase in the random stage sample, the results of the item selection methods on the accuracy of *b* parameter for the 2-PL model are not consistent as in the above-mentioned results. The differences between these sample sizes are more apparent. For MFI and small calibration sample size (*N* = 250), the accuracy of parameter estimation is reduced as the sample size of the random phase increases. When the parameter estimation method is JML, BDOD and DVOD showed the same trend for *N* = 250 and 500, respectively. In contrast, DVOD behaved the opposite way when *N* = 500 for OEM and when *N* = 1000 for JML. In other conditions, RMSE fluctuations were seen for different pretest item selection methods at different sample levels.

To compare the effect of the random phase sample size via parameter estimation methods, RMSE values of these for other conditions crossed by the pretest item selection methods and calibration sample size of per pretest item (*N* = 250, 500, and 1000) in Tables 3 and 4 are plotted in Figure 16 for the 1-PL model and *b* parameter, Figure 17 for the 2-PL model and *a* parameter, and Figure 18 for the 2-PL model and *b* parameter.

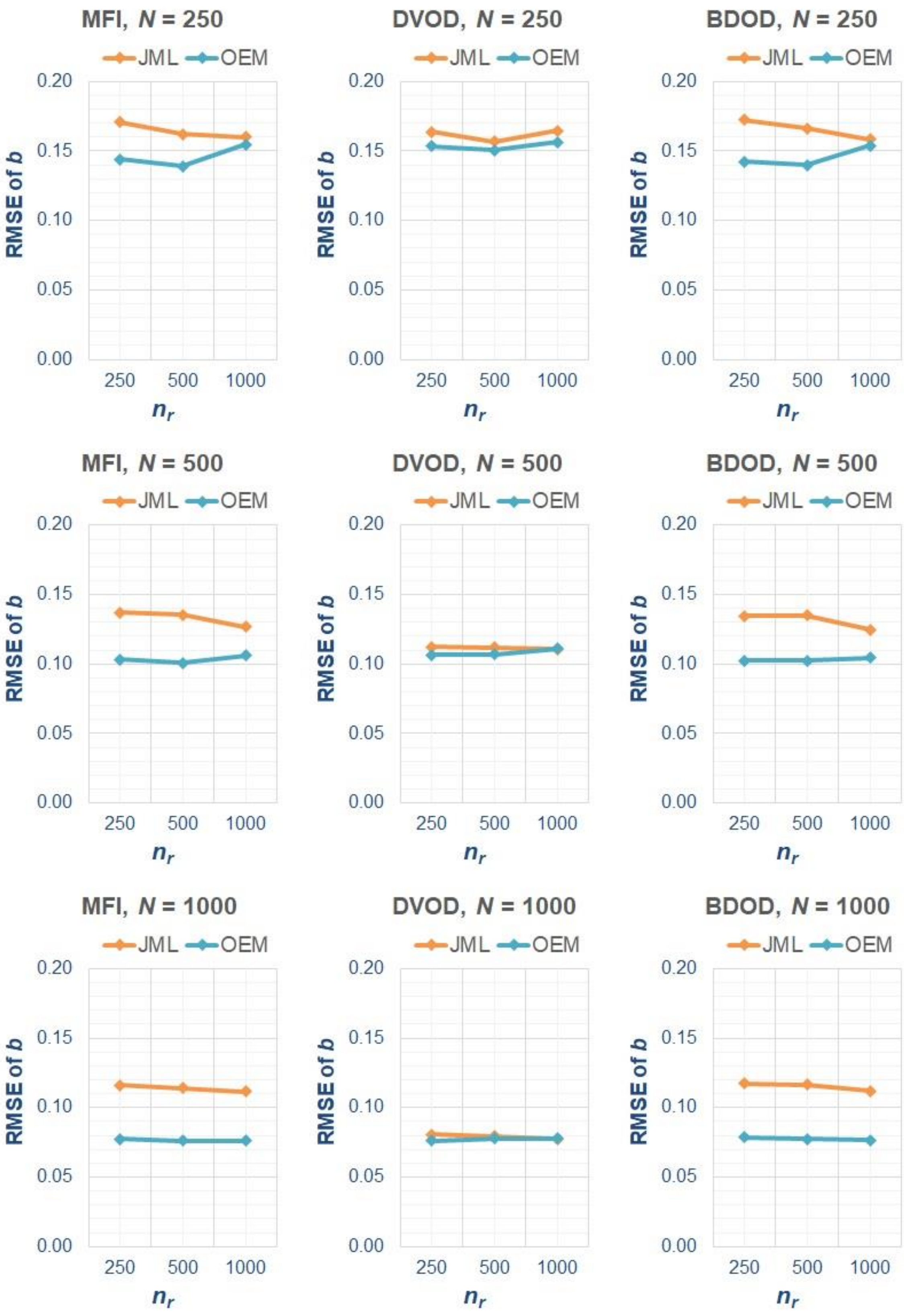


Figure 16. RMSE (*b* parameter) of the parameter estimation method for sample size of the random calibration stage under 1-PL model and different conditions

The effect of the random phase sample size via the calibration methods on parameter accuracy for the 1-PL model and *b* parameter is comparable as can be seen from Table 3 and Figure 16. JML produced more precise parameter estimation with an increasing sample size of the random stage in almost all conditions. Even the largest decrease (from 0.1725 to 0.1584; see right-top in Figure 16) in RMSE is for JML when the sample size increases from n_r = 250 to n_r = 1000 and BDOD used as pretest item selection method. OEM showed a tendency to decrease and then rise in most conditions for *N* = 250 and 500 when the random phase samples increased from 250 to 500 and from 500 to 1000. For *N* = 1000, the changes in different n_r levels were negligible in all item selection methods.

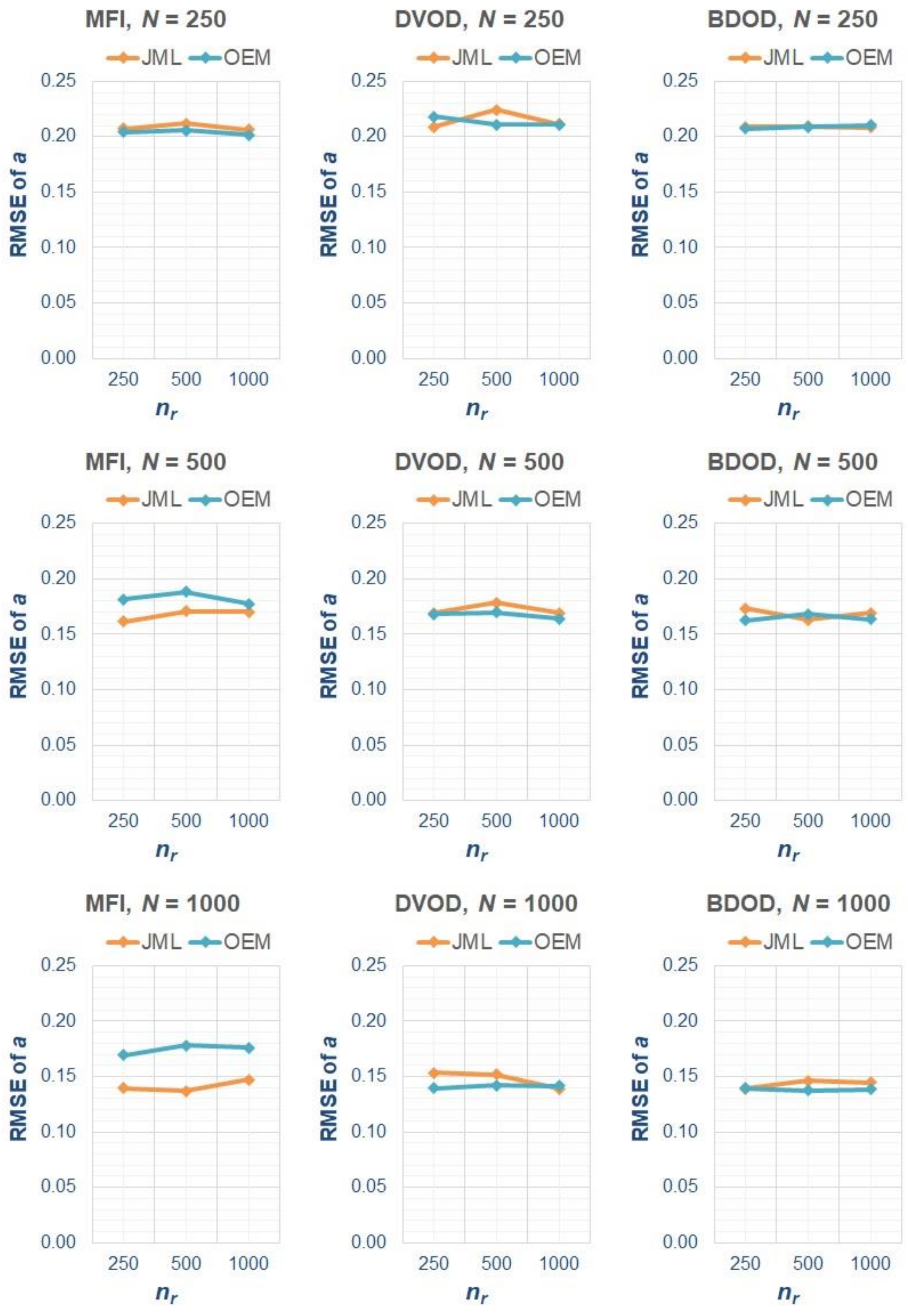


Figure 17. RMSE (*a* parameter) of the parameter estimation method for sample size of the random calibration stage under 2-PL model and different conditions

From Table 4 and Figure 17, with the increasing of random phase sample size, the performance of the calibration methods for the 2-PL model and *a* parameter is irregular. In most conditions; the fluctuations were detected in both JML and OEM, where there was an increase from n_r = 250 to n_r = 500 followed by a decline from n_r = 500 to n_r = 1000. However, these differences between different sample sizes of the random stage for the parameter estimation methods are minor levels. Apart from that, as can be seen from the overlapping two lines on the graph on the right-top in Figure 17, the performances of the two methods for BDOD are extremely close with the increasing of random phase sample size when the calibration sample size is small (*N* = 250).

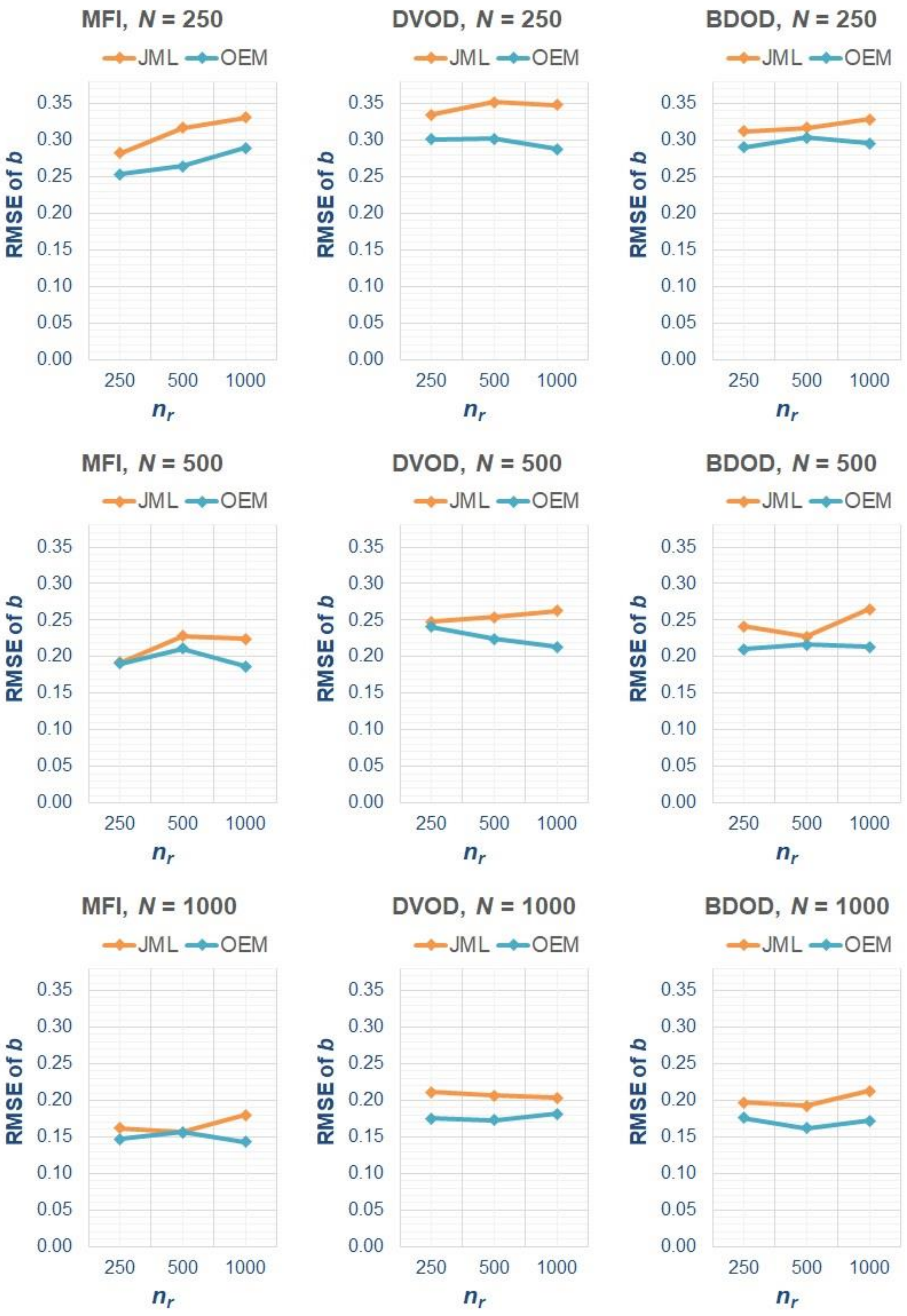


Figure 18. RMSE (*b* parameter) of the parameter estimation method for sample size of the random calibration stage under 2-PL model and different conditions

As can be seen from Table 4 and Figure 18, the parameter estimation methods have different tendencies under the same conditions as the increasing sample size of the random stage from small (n_r = 250) to large (n_r = 1000). They have similar trend for MFI when N = 250 (continuous increase) and N = 500 (first increase then decline), for DVOD when N = 250 (first increase then decline), and for BDOD when N = 1000 (first decrease then rise). However, JML and OEM performed in the opposite way for MFI when N = 1000 (first decrease then rise and first increase then decline, respectively), for DVOD when N = 500 (continuous increase and continuous decline, respectively), and for BDOD when N = 250 (first decrease then rise and first increase then decline, respectively) and N = 500 (first decrease then rise and first increase then decline, respectively). Moreover, the greatest difference (from 0.2825 to 0.3304) between the random phase sample sizes (from 250 to 1000) was observed when MFI was used as the parameter estimation method and N = 250

In summary, the effect of the increasing sample size of the random stage has changed according to IRT model, item parameters, pretest item selection methods and parameter estimation methods. However, these effects (i.e., continuous increase, first decrease then rise) cannot be generalized in terms of these variables. Sample size of the random calibration stage was considered as a variable only in He et al. (2020) study while it was fixed to 150 for all pretest items in He (2015) and was approximately set 100 for each pretest items in Zheng (2014), Zheng (2016), and Zheng and Chang (2017) studies. He et al. (2020) examined the effect of the different sample sizes of random calibration stage (100, 200, 300, 400, and 500) on parameter estimation accuracy by calculating the relative efficiency as evaluation criteria. In their study, only OEM parameter estimation method was used and BDOD performance is better for larger sample sizes whereas Excellence Degree (ED) criterion works well in small random samples. For BDOD and the 2-PL model, this finding was observed only when the calibration sample size was large (N = 1000) in this study. He et al. (2020) explained this by the fact that DVOD is more sensitive than ED to more accurate initial parameter estimation with larger random stage sample size. However, ED compensated for this disadvantage with its adaptive stage performance. In this study, the irregularities of different pretest item selection methods and parameter estimation methods can be explained in this way.

Comparison Calibration Sample Size of Per Pretest Item

To compare the effect of the calibration sample size of per pretest item (250, 500, 1000) on precision in parameter estimation, average RMSE values were used. This effect was investigated separately for item selection methods and parameter estimation methods, respectively. As explained above, no tables are included in this section. RMSE values of pretest item selection methods at these sample size for other conditions crossed by the parameter estimation methods and the sample size of random stage (n_r = 250, 500, and 1000) in Tables 3 and 4 are graphed in Figure 19 for the 1-PL model and *b* parameter, Figure 20 for the 2-PL model and *a* parameter, and Figure 21 for the 2-PL model and *b* parameter.

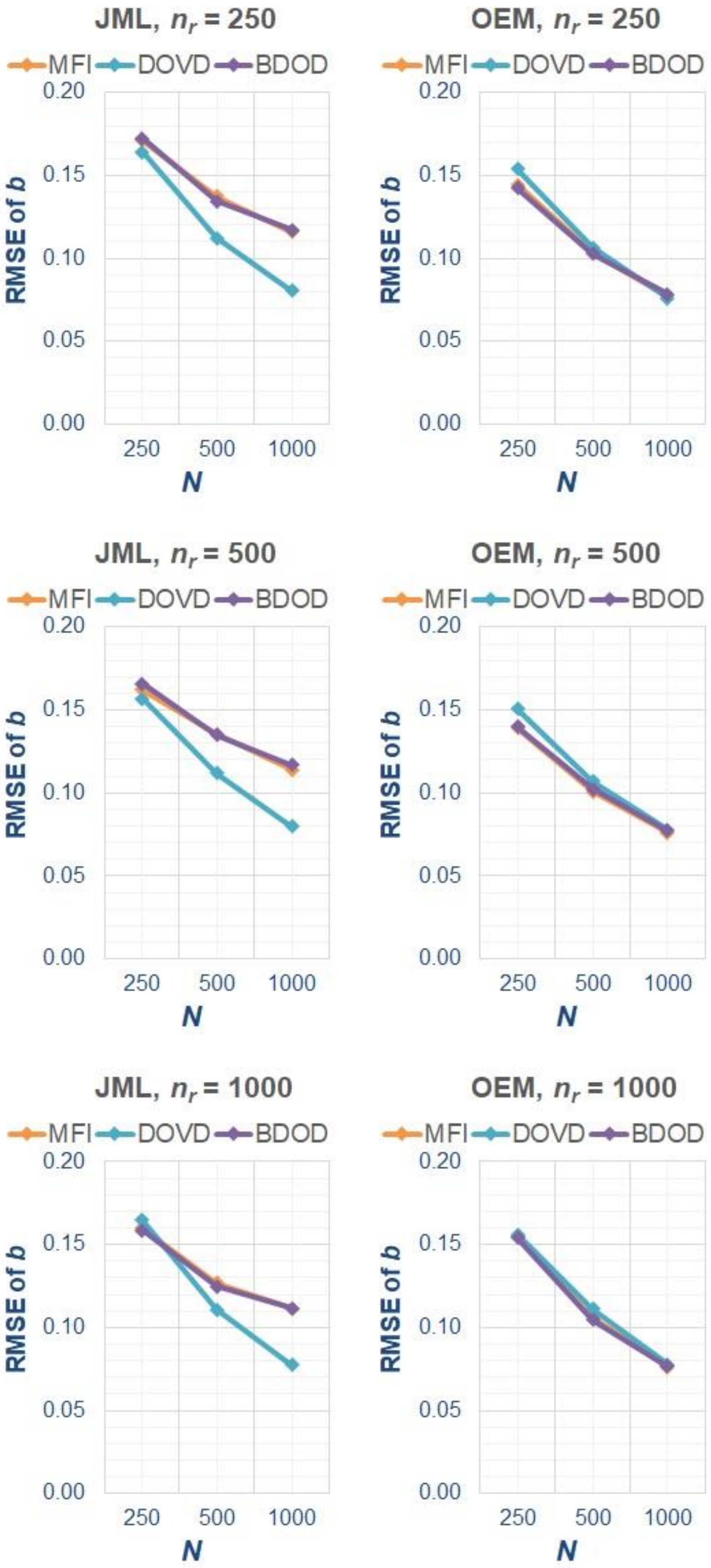


Figure 19. RMSE (*b* parameter) of the item selection method for calibration sample size under 1-PL model and different conditions

The effect of the calibration sample size via the pretest item selection methods on parameter accuracy for the 1-PL model and *b* parameter is similar in all simulated conditions, as can be seen from Table 3 and Figure 19. As expected, when the calibration sample increased from 250 to 1000, the RMSE value for all methods dropped, in other words, the parameter accuracy improved. This decrease was sharp for DVOD when the parameter estimation method is JML (see the left side in Figure 19), but it was considerable for other conditions.

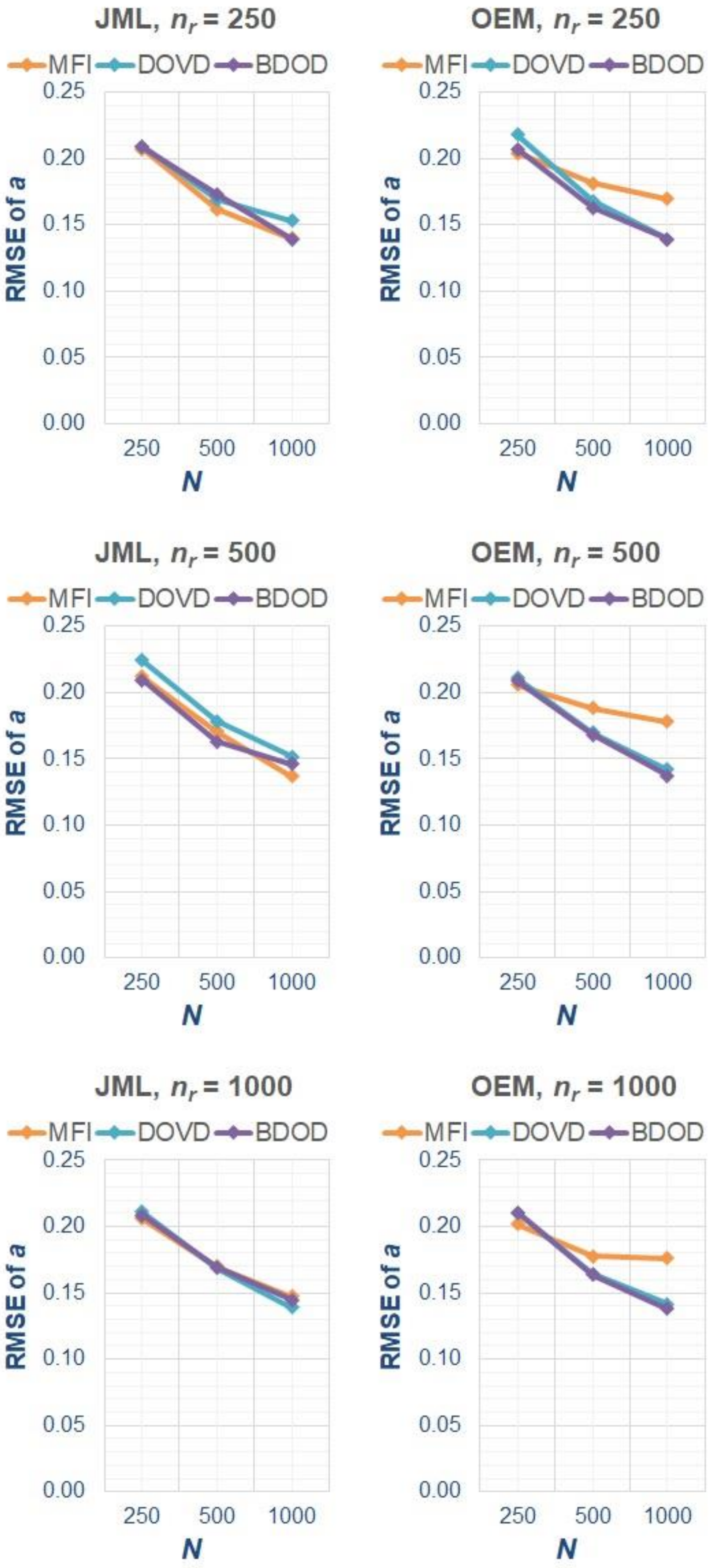


Figure 20. RMSE (*a* parameter) of the item selection method for calibration sample size under 2-PL model and different conditions

According to Table 4 and Figure 20, the pretest item selection methods for the 2-PL model and *a* parameter under all conditions consistently yielded smaller RMSE values with the increasing of calibration sample size under the parameter estimation methods. This means the larger sample size generated better parameter recovery. However, the variation between calibration sample sizes is at different levels. As the calibration sample increased from 250 to 1000, RMSE value went down gradually when OEM used as the parameter estimation method for MFI (see the right side in Figure 20) whereas it decreased substantially under other conditions.

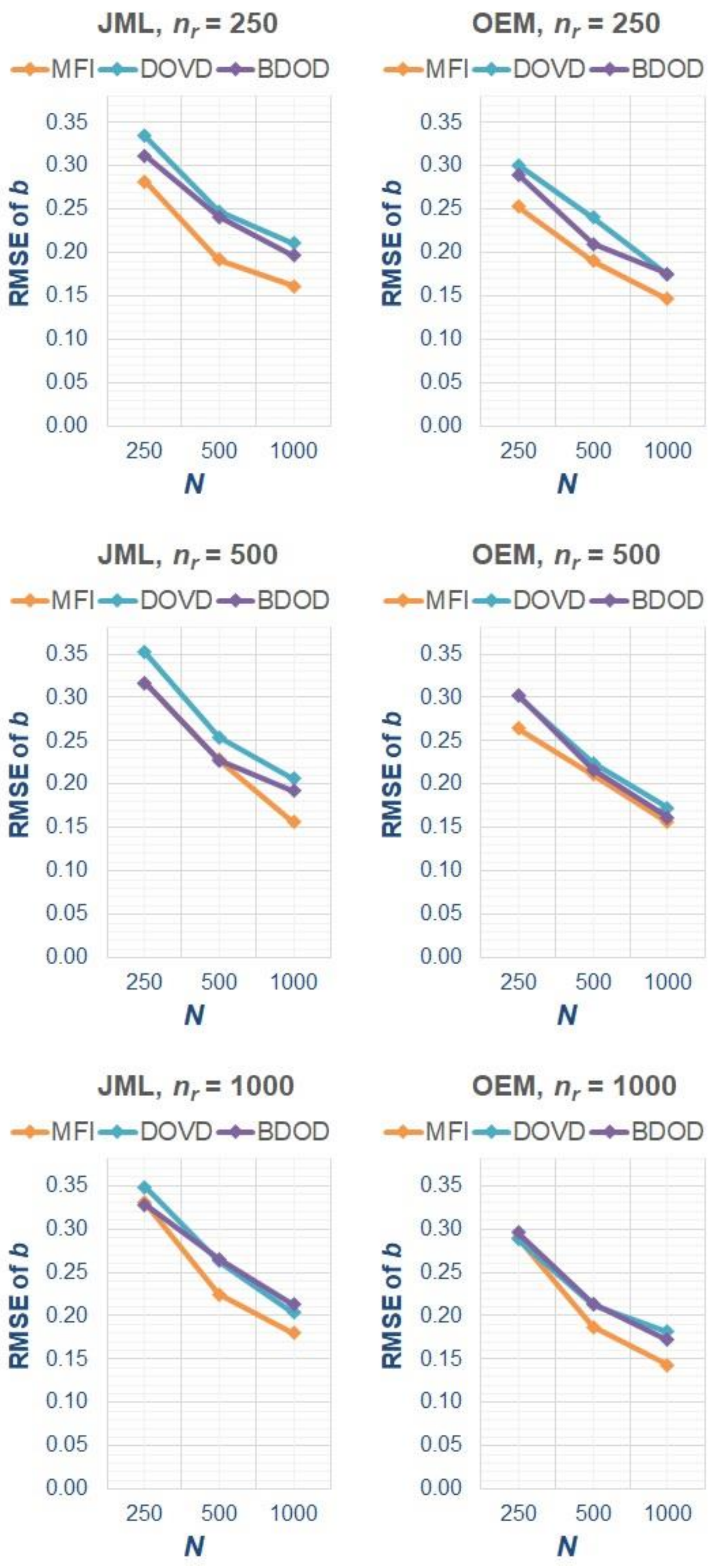


Figure 21. RMSE (*b* parameter) of the item selection method for calibration sample size under 2-PL model and different conditions

As Table 4 and Figure 21 show, the results of the item selection methods on the accuracy of *b* parameter for the 2-PL model improved from small calibration sample (*N* = 250) to large sample (*N* = 250) in all simulated conditions. In terms of RMSE values, small calibration samples produced the worst performance whereas large samples had the best performance. Moreover, these changes are more rapidly - especially from *N* = 250 to *N* = 500 - than the discrimination parameters.

To compare the effect of the calibration sample sizes of per pretest item size via parameter estimation methods, RMSE values of these for other conditions crossed by the pretest item selection methods and the sample sizes of random phase (n_r = 250, 500, and 1000) in Tables 3 and 4 are plotted in Figure 22 for 1-PL model and *b* parameter, Figure 23 for 2-PL model and *a* parameter, and Figure 24 for 2-PL model and *b* parameter.

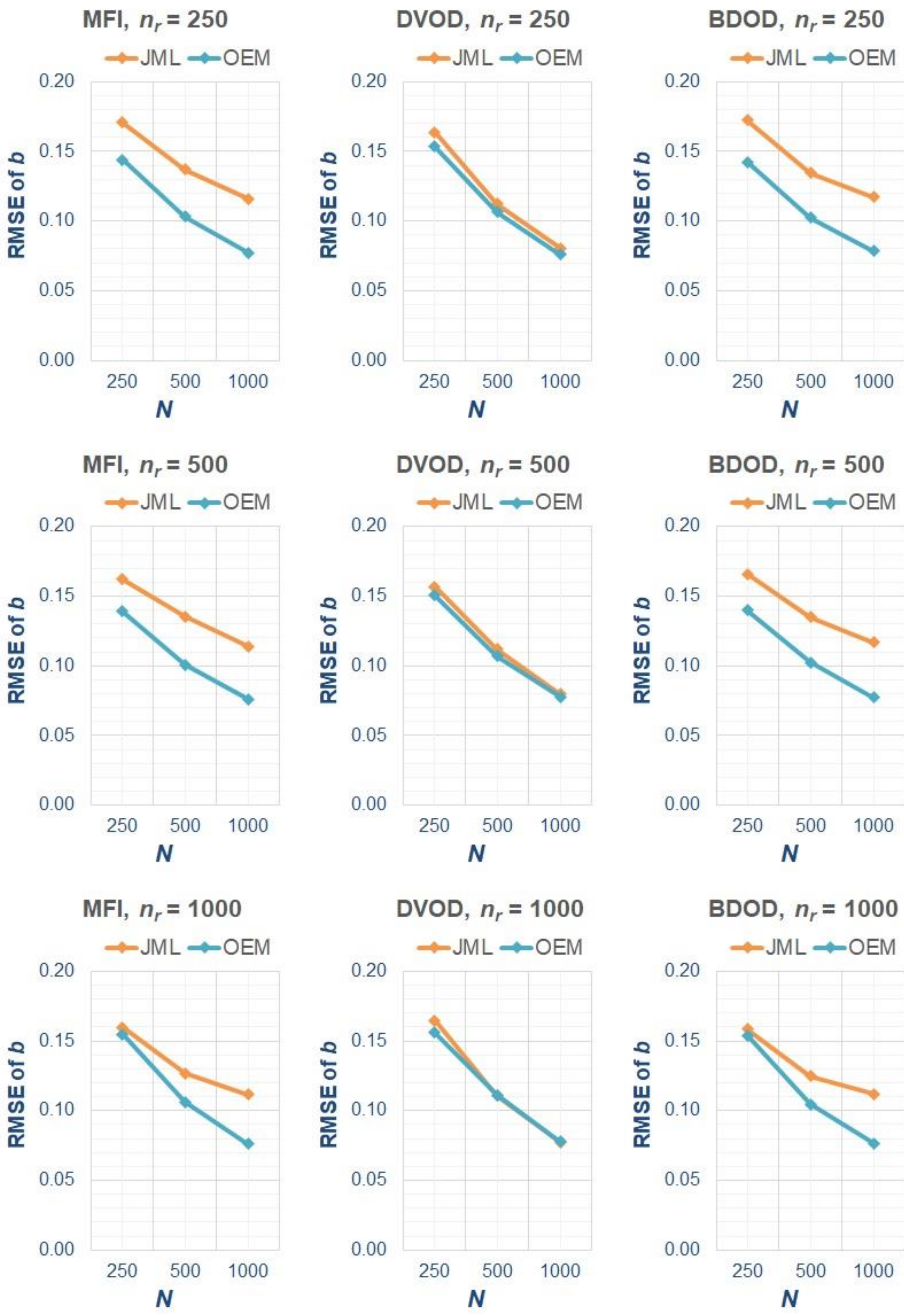


Figure 22. RMSE (*b* parameter) of the parameter estimation method for calibration sample size under 1-PL model and different conditions

The effect of the calibration sample size via the calibration methods on parameter accuracy for the 1-PL model and *b* parameter is similar as can be seen from Table 3 and Figure 22. Both JML and OEM produced more accurate parameter estimation as the number of samples increased. At different calibration sizes, For DVOD, the performance of these methods is almost the same whereas OEM performed better than JML for MFI and DVOD except when n_r = 1000 and *N* = 250. However, MFI and DVOD showed similar decreases when the calibration sample size increased -except again n_r = 1000 - (see the left-top, the left-middle, the right-top and the right-middle in Figure 22). For n_r = 1000, the performance variation between JML and OEM has been going up in favour of OEMs with an increase from *N* = 250 to *N* = 1000.

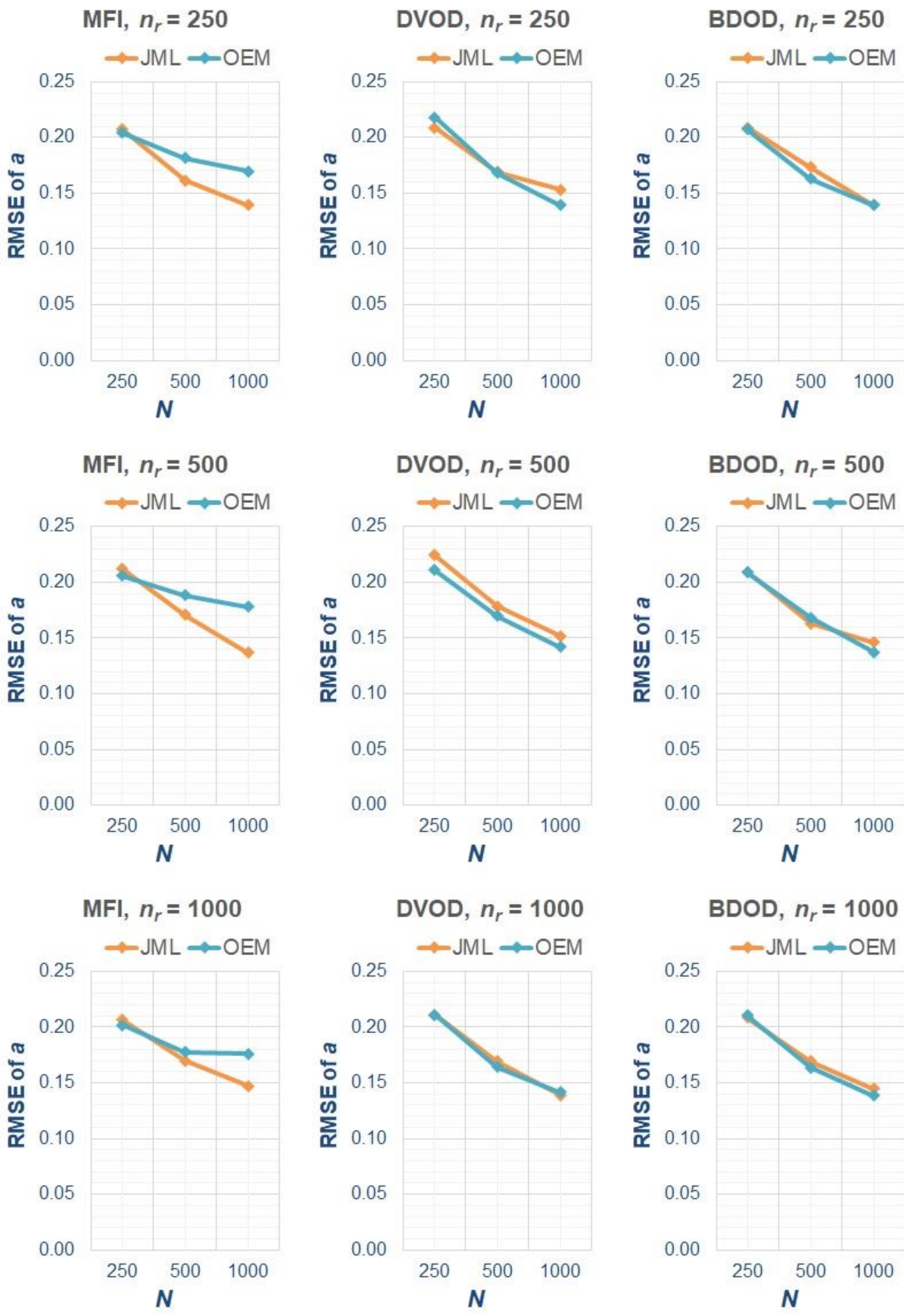


Figure 23. RMSE (*a* parameter) of the parameter estimation method for calibration sample size under 2-PL model and different conditions

From Table 4 and Figure 23, with the increasing of calibration sample size, the parameter estimation methods for the 2-PL model and *a* parameter performed very similarly according to the pretest item selection method. For DVOD and BDOD, these calibration methods generated very close RMSE values at different calibration sizes. For MFI, JML yielded better parameter recovery than OEM as increasing the calibration sample size. Especially, the differentiation between JML and OEM was maximum under this condition when N = 1000.

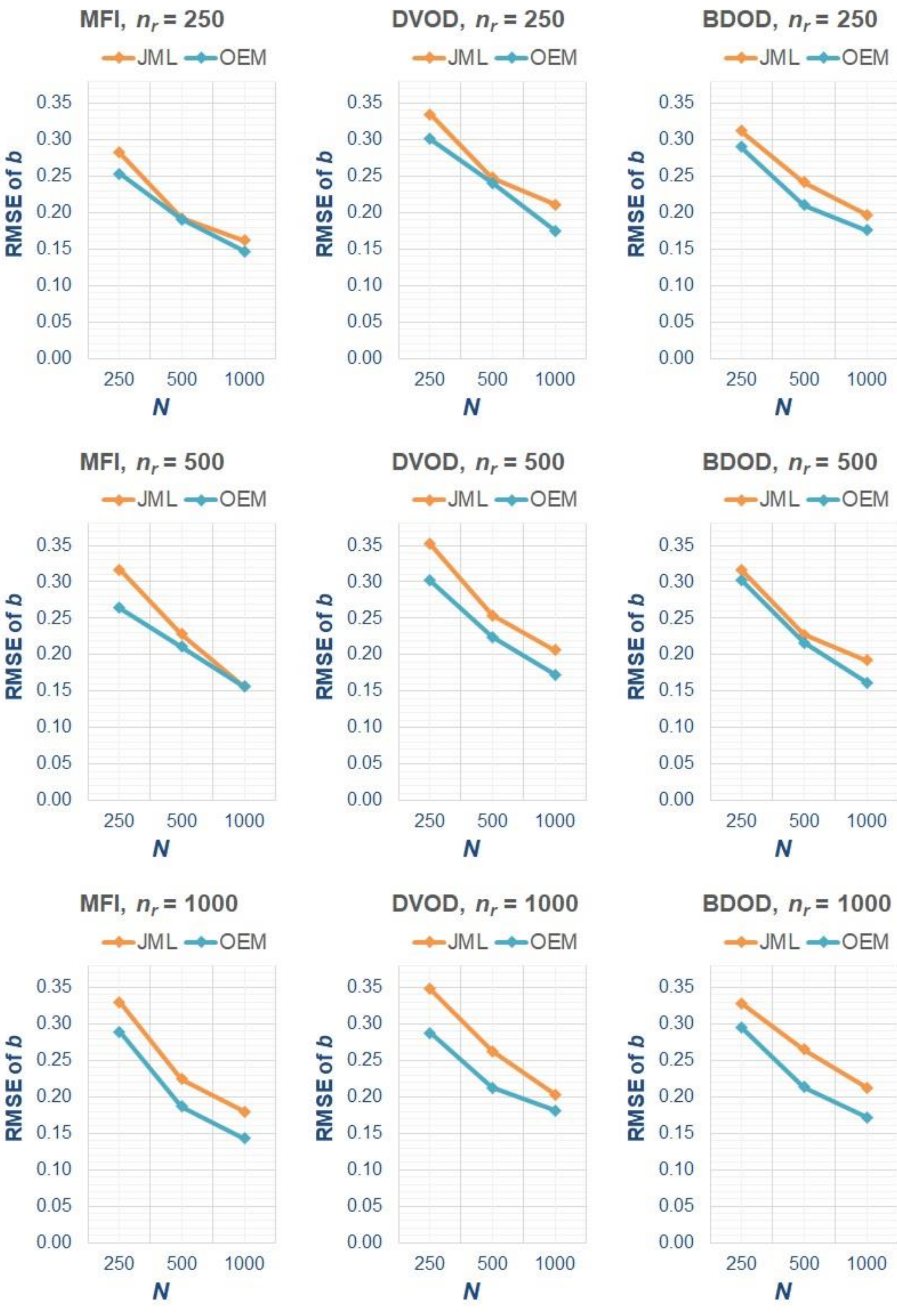


Figure 24. RMSE (*b* parameter) of the parameter estimation method for sample calibration sample size under 2-PL model and different conditions

As can be seen from the Table 4 and Figure 24, the parameter estimation methods tend to improve parameter accuracy as increasing the calibration sample size from small (*N* = 250) to large (*N* = 1000). For MFI and DVOD, the performances of JML and OEM are very close when n_r = 250 and *N* = 500 (see the left-top and the center-top in Figure 24). In most other conditions, the parameter estimation methods showed a similar reduction trend; the substantial from n_r = 250 to n_r = 500 then moderate from n_r = 500 to n_r = 1000.

Ali and Chang (2014), Ban et al. (2001), He (2015), He et al. (2017), He et al. (2020) and Kingsbury (2009) tested the effect of calibration sample size on the accuracy of parameter estimation. Although some fluctuations are seen with the arising sample size in Ali and Chang's (2014) study, a larger sample size improves parameter recovery for all parameters, the pretest item selection method and the parameter estimation method in all studies, as expected. This finding indicates that the problem of the classical version of JML that the failure to improve parameter accuracy as the number of responses is not encountered in the context of online calibration. This can be explained by the fact that JML can be converged more easily due to This can be explained by the fact that JML can be converged more easily due to its single step application.

Chapter 5
Conclusion and Suggestions

In this section, the results of this study are summarized, and then suggestions are made for both practical applications of the study and future research.

Conclusion

The purpose of this study is to investigate the effect of the online calibration elements (pretest item selection methods, the parameter estimation methods, the sample size of the random calibration stage and the calibration sample size of per pretest item) on the accuracy of pretest item parameter. In addition, this study also aimed to use JML as a parameter estimation method to the online calibration procedure and assess this method's feasibility. In line with this purpose, a simulation study was carried out. For the 108 conditions formed by crossing the IRT models (1-PL and 2-PL), pretest item selection methods (MFI, D-optimal value design, and Bayesian D-optimal design), parameter estimation methods (JML, and, OEM), the sample size of random calibration stage (250, 500, and, 1000) and the total number of responses (250, 500, and, 1000), the parameters of the pretest items were estimated and then the RMSE and bias values were calculated. The findings obtained from the simulation study are as follows.

- Compared to the pretest item selection methods, it was seen that the results for the 1-PL model and *b* parameter differed according to parameter estimation methods. D-optimal value design is the best method for JML, while Maximum Fisher Information and Bayesian D-optimal design are the best and close performance method for OEM. For the 2-PL model and *a* parameter, the effectiveness of the methods also depends on the parameter estimation method. When the parameter estimation method is JML, it is seen that MFI performs a little better although a predominant method does not stand out among them. For OEM, Bayesian D-optimal design has been the best pretest item selection method by producing the lowest RMSE in most conditions. The worst method is D-optimal value design for small calibration sample size (*N* = 250) and MFI for medium and large calibration sample sizes (*N* = 500 and 1000). Finally, for the 2-PL model and *b* parameter, MFI is seemed to be the best choice among the pretest item selection methods, regardless of the calibration method.

- When the item selection methods were compared in terms of the cumulative sample size, it was seen that the use of D-optimal value design as the item selection method was inclined to early retirement for both IRT models and parameter estimation methods. This causes to the priority administration of pretest items which are favorite in terms of the D-optimal value. Bayesian D-optimal design behaved similarly for the 2-PL model. In addition, MFI and Bayesian D-optimal design for the 1-PL model and Bayesian D-optimal design for the 2-PL model have completed the calibration process with more average simulated examinee.
- Comparing the performance of the parameter estimation methods, OEM produced a lower RMSE value for the 1-PL model and parameter b, regardless of the pretest item selection method. When the pretest item selection method is D-optimal value design, JML performed close to OEM. For the 2-PL model and *a* parameter, the results of both the parameter estimation methods are neck and neck. Specifically, JML worked slightly better than OEM at medium and large calibration sizes (*N* = 500 and 1000) when MFI was used. For the 2-PL model and *b* parameter, OEM has successfully estimated the most precise item parameters.
- Assessing the effectiveness of the parameter estimation method at different levels, it was observed that they showed extreme deviations for the easiest, most difficult and more discriminating items. Apart from this, *a* parameter is underestimated under all conditions.
- When the effect of the sample size of random calibration stage in terms of parameter estimation accuracy was examined, the increase of it caused different effects (such as a decrease after an increase, a continuous decrease, a continuous increase, an increase after a decrease) in IRT model, item selection method, parameter estimation method and calibration sample sizes. Therefore, it is not possible to observe a specific trend for these tested factors.
- Lastly, the effect of increasing the calibration sample size is similar for all pretest item selection methods and all parameter estimation methods, although it

occurs at different levels. The parameter accuracy gets higher as the calibration sample size increases.

Suggestions

In this section, suggestions for practice based on the research results and future research are presented below.

Suggestions for practice based on the research results

1. MFI produced more efficient results as the pretest item selection method when both OEM and JML were used for the 1-PL model. According to these results, the use of MFI in online calibration applications for the 1-PL model will be beneficial thanks to its simple and easily applicable structure.
2. For the 2-PL model, MFI and OEM methods may be preferred as a pretest item selection method and parameter estimation method, respectively. However, Bayesian D-optimal design may be an option for the 2-PL model since MFI lagged behind it in terms of the precision of the discrimination parameter.
3. D-optimal value design method may be preferred to complete the calibration process with fewer examinees in cases where it is not a problem to apply the pretest items with the highest optimal D value without considering the ability level of the examinee.
4. In similar studies, it may be recommended to use a larger sample size for pretest item calibration as it increases the parameter accuracy.

Suggestions for future research

1. In this research, JML as a new pretest parameter estimation method was compared with OEM only. In future research, the effect of existing calibration methods can be compared with JML.
2. In this study, JML was found to be ineffective in most cases. In future research, the effectiveness of JML can be improved by employing different statistical correction methods.

3. In this study, the parameter update sample size is set to 10 new responses for each pretest item. In future research, its effect on the accuracy of parameter estimation can be tested by using larger samples to reduce the calibration time.

4. In this research, the parameter estimation of a pretest item is completed when it is administrated to the specified number of samples. The impact of different termination rules (for example, based on the standard error of parameter estimation, depending on the stability of the changes in the estimation of the pretest item parameter) can be examined in future research.

5. In this study, item parameters in both operational item bank and pretest item bank and the ability parameter of examinee were generated to mimic the real-life situation. In future studies, post-hoc or hybrid simulations can be carried out using real items and ability parameters. By going one step further, the effectiveness of online calibration components can be tested in practice by demonstrating live online calibration application.

6. In this study, CAT components such as first and next item selection method and parameter estimation method are fixed in CAT phase. Different methods of these components can be selected in future studies. For example, different operational items may be selected depending on the different item selection method, and therefore the results of the OEM method may be influenced by the inclusion of these items parameter in the calculation.

7. In this study, although the running time were logged, it was not taken into consideration because the simulation studies were performed on different computers and virtual machines with different hardware. In future studies, it can be used as a variable and thus more information about the JML method can be obtained.

.

References

Aybek, E. C., & Demirtasli, R. N. (2017). Computerized adaptive test (CAT) applications and item response theory models for polytomous items. *International Journal of Research in Education and Science, 3(*2), 475-487. https://doi.org/10.21890/ijres.327907

Ali, U. S., & Chang, H-H. (2014). An Item-Driven Adaptive Design for Calibrating Pretest Items. *ETS Research Report Series, 2014*(2), 1-12. https://doi.org/10.1002/ets2.12044

Amazon Web Services. (2019). Amazon Elastic Compute Cloud (EC2). Retrieved from https://aws.amazon.com/ec2/

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis* (2nd ed.). New York, NY: Wiley. https://doi.org/10.2307/2531310

Ban, J. C., Hanson, B. A., Wang, T., Yi, Q., & Harris, D. J. (2001). A Comparative Study of On-line Pretest Item-Calibration/Scaling Methods in Computerized Adaptive Testing. *Journal of Educational Measurement, 38*(3), 191-212. https://doi.org/10.1111/j.1745-3984.2001.tb01123.x

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Marcel Dekker. https://doi.org/10.1201/9781482276725

Berger, M. P. (1992). Sequential sampling designs for the two-parameter item response theory model. *Psychometrika*, *57*(4), 521-538. https://doi.org/10.1007/BF02294418

Berger, M. P. (1994). D-optimal sequential sampling designs for item response theory models. *Journal of Educational Statistics*, *19*(1), 43-56. https://doi.org/10.3102/10769986019001043

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Bock, R. D. & Mislevy, R. J. (1988). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement*, *6*, 431–444. https://doi.org/10.1177/014662168200600405

Brent, R. (1973). *Algorithms for minimization without derivatives.* Prentice-Hall. ISBN: 0-13-022335-2

Bulut, O., & Kan, A. (2012). Application of computerized adaptive testing to entrance examination for graduate studies in Turkey. *Eurasian Journal of Educational Research, 49*, 61-80.

Buyske, S. G. (1998). Optimal design for item calibration in computerized adaptive testing: the 2PL case. *New developments and applications in experimental design, 34,* 115-125. https://doi.org/10.1214/lnms/1215456191

Chang, Y. C. I., & Lu, H. Y. (2010). Online calibration via variable length computerized adaptive testing. *Psychometrika, 75*(1), 140-157. https://doi.org/10.1007/s11336-009-9133-0

Chang, H.-H. & Ying, Z. (1996). A global information approach to computerized adaptive testing. *Applied Psychological Measurement, 20*, 213–229. https://doi.org/10.1177/014662169602000303

Chen, P. (2017). A comparative study of online item calibration methods in multidimensional computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 42*(5), 559-590. https://doi.org/10.3102/1076998617695098

Chen, P., Xin, T., Wang, C., & Chang, H. H. (2012). Online calibration methods for the DINA model with independent attributes in CD-CAT. *Psychometrika, 77*(2), 201-222. https://doi.org/10.1007/s11336-012-9255-7

Chen, P., & Wang, C. (2016). A new online calibration method for multidimensional computerized adaptive testing. *Psychometrika, 81*(3), 674-701. https://doi.org/10.1007/s11336-015-9482-9

Chen, P., Wang, C., Xin, T., & Chang, H. H. (2017). Developing new online calibration methods for multidimensional computerized adaptive testing. *British Journal of Mathematical and Statistical Psychology*, *70*(1), 81-117. https://doi.org/10.1111/bmsp.12083

Choi, S. W. (2009). Firestar: Computerized adaptive testing simulation program for polytomous item response theory models. *Applied Psychological Measurement, 33*(8), 644-645. https://doi.org/10.1177/0146621608329892

Colvin, K. F. (2014). Effect of automatic item generation on ability estimates in a multistage test. (Unpublished Doctoral Dissertation). Amherst, MA.: University of Massachusetts.

Davey, T., Nering, M. L., & Thompson, T. (1997). *Realistic simulation of item response data* (ACT Research Report Series No. 97-4). Iowa City, IA: ACT.

de Gruijter, D. N. (1990). A note on the bias of UCON item parameter estimation in the Rasch model. *Journal of Educational Measurement, 27*(3), 285-288. https://doi.org/10.1111/j.1745-3984.1990.tb00749.x

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistic model. *Applied Psychological Measurement, 13*, 77-90. https://doi.org/10.1177/014662168901300108

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York: Springer. https://doi.org/10.1007/978-1-4614-6868-4

Eddelbuettel D, François R, Allaire J, Ushey K, Kou Q, Russel N, Chambers J, Bates D (2018). *Rcpp: Seamless R and C++ Integration.* R package version 0.12.15, URL http: //CRAN.R-Project.org/package=Rcpp.

Eddelbuettel, D., François, R., Bates, D., & Ni, B. (2019). RcppArmadillo: Rcpp integration for Armadillo templated linear algebra library. *R package version 0.9*, *800*(0).

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Feinberg, R. A., & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36–49. https://doi.org/10.1111/emip.12111

Flaugher, R. (2000). Item pools. In H. Wainer, Dorans, N. J., R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing: A primer* (pp. 171-199). Mahwah, NJ: Erlbaum.

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling, 60*(3), 327-346.

Guo, R. (2016). *Item parameter drift and online calibration.* (Doctoral dissertation). University of Illinois, Urbana-Champaign.

Han, K. T. (2012). SimulCAT: Windows software for simulating computerized adaptive test administration. *Applied Psychological Measurement, 36*(1), 64-66. https://doi.org/10.1177/0146621611414407

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer. https://doi.org/10.1007/978-94-017-1988-9

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage Publications, Inc.

He, W. (2015). CAT Field-Test Item Calibration Sample Size: How Large is Large under the Rasch Model?. *Global Journal of Human-Social Science Research. 15*(1).

He, Y., Chen, P., & Li, Y. (2020). New efficient and practicable adaptive designs for calibrating items online. *Applied Psychological Measurement, 44*(1), 3-16. https://doi.org/10.1177/0146621618824854

He, Y., Chen, P., Li, Y., & Zhang, S. (2017). A New Online Calibration Method Based on Lord's Bias-Correction. *Applied Psychological Measurement, 41*(6), 456-471. https://doi.org/10.1177/0146621617697958

Holland, P. W. (1990). On the sampling theory foundations of item response theory models. *Psychometrika, 55*, 577–602. https://doi.org/10.1007/BF02294609

Hsu, Y., Thompson, T. D., & Chen, W. H. (1998). *CAT item calibration.* In annual meeting of the National Council on Measurement in Education, San Diego.

Jones, D. H., & Jin, Z. (1994). Optimal sequential designs for on-line item estimation. *Psychometrika*, *59*(1), 59-75.

Kingsbury, G. (2009). *Adaptive item calibration: A process for estimating item parameters within a computerized adaptive test.* Paper presented at the Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing.

Linacre, J. M. (2000). Computer-Adaptive Testing : A Methodology whose time has come. In S. Chae, U. Kang, E. Jeon, & J. M. Linacre (Eds.), *Development of*

Computerized Middle School Achievement Test (pp. 1–58). Seoul: Komesa Press.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F.M. (1980a). Some how and which for practical tailored testing. In L.J.T. van der Kamp, W.F. Langerak et D.N.M. de Gruijter (éds): *Psychometrics for educational debates.* New York: John Wiley and Sons.

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

Lord, F. M. (1983). Unbiased estimators of ability parameters, of their variance, and of their parallel-forms reliability. *Psychometrika, 48*(2), 233-245. https://doi.org/10.1007/BF02294018

Lu HY (2014) Application of Optimal Designs to Item Calibration. *Plos One 9*(9): e106747. https://doi.org/10.1371/journal.pone.0106747

Luecht, R. M. & Sireci, S. G. (2011). A review of models for computer-based testing. *Research Report RR-2011-12*. New York: The College Board.

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerized adaptive and multistage testing with R: Using packages catR and mstR*. Springer. https://doi.org/10.1007/978-3-319-69218-0

Magis, D., & Raîche, G. (2012). Random generation of response patterns under computerized adaptive testing with the R package catR. *Journal of Statistical Software*, *48*(8), 1-31. https://doi.org/10.18637/jss.v048.i08

Magis, D., Raiche, G., & Barrada,J. R. (2018). Package 'catR'. *R package, version*, *3*.

Makransky, G., & Glas, C. A. (2014). An automatic online calibration design in adaptive testing. *Journal of Applied Testing Technology, 11*(1), 1-20.

Mead, A. D., & Drasgow, F. (1993). Equivalence of computerized and paper-and-pencil cognitive ability tests: *A meta-analysis. Psychological Bulletin, 114,* 449–458. https://doi.org/10.1037/0033-2909.114.3.449

Microsoft Azure. (2019). Azure Virtual Machines. Retrieved from https://azure.microsoft.com/en-us/services/virtual-machines/

Mills, C. N., & Stocking, M. L. (1995). Practical Issues in Large-Scale High-Stakes Computerized Adaptive Testing. *ETS Research Report Series, 1995*(2), i-28. https://doi.org/10.1002/j.2333-8504.1995.tb01658.x

Mislevy, R. J. (1984). Estimating latent distributions. *Psychometrika*, *49*(3), 359-381. https://doi.org/10.1007/BF02306026

Natesan, P., Nandakumar, R., Minka, T., & Rubright, J. D. (2016). Bayesian prior choice in IRT estimation using MCMC and variational Bayes. *Frontiers in psychology, 7*, 1422. https://doi.org/10.3389/fpsyg.2016.01422

Nydick, S. W. (2014). catIrt: An R package for simulating IRT-based computerized adaptive tests. *R package, version 0.5-0*.

Owen, R. J. (1975). A Bayesian sequential procedure for quantal response in the context of adaptive mental testing. *Journal of the American Statistical Association*, *70*, 351–356. https://doi.org/10.1080/01621459.1975.10479871

Özberk, E. H., & Gelbal, S. (2017). A Comparison of Multidimensional Item Selection Methods in Simple and Complex Test Designs. *Journal of Measurement and Evaluation in Education and Psychology, 8*(1), 34-46. https://doi.org/10.21031/epod.286956

Parshall, C. G. (1998). *Item development andpretesting in a computer-based testing environment.* Paper presented at the colloquium Computer-Based Testing: Building the Foundation for Future Assessments, Philadelphia, PA.21.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Reckase, M. D. (1989). Adaptive testing: The evolution of a good idea. *Educational Measurement: Issues and Practice, 8*(3), 11-15. https://doi.org/10.1111/j.1745-3992.1989.tb00326.x

Ren, H., van der Linden, W. J., & Diao, Q. (2017). Continuous online item calibration: Parameter recovery and item utilization. *Psychometrika*, *82*(2), 498-522. https://doi.org/10.1007/s11336-017-9553-1

Revuelta, J., & Ponsoda, V. (1998). A comparison of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 311–327. https://doi.org/10.1111/j.1745-3984.1998.tb00541.x

Rubinstein, R. Y., & Kroese, D. P. (2017). *Simulation and the Monte Carlo Method* (3rd ed.). Hoboken, NJ: Wiley. https://doi.org/10.1002/9781118631980

Rudner, L. M. (1998). An on-line, interactive computer adaptive testing mini tutorial. Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores* (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf

Segall, D. O. (2004). A sharing item response theory model for computerized adaptive testing. *Journal of Educational and Behavioral Statistics, 29*, 439–460. https://doi.org/10.3102/10769986029004439

Stocking, M. L. (1988). Scale drift in on-line calibration. *ETS Research Report Series, 1988*(1), i-122. https://doi.org/10.1002/j.2330-8516.1988.tb00284.x

Stout, W., Ackerman, T., Bolt, D., Froelich, A. G., & Heck, D. (2003). On the Use of Collateral Item Response Information To Improve Pretest Item Calibration. LSAC Research Report Series.

Team, R. C. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Thissen, D., & Mislevy, R. J. (2000). Testing algorithms. In H. Wainer, (Ed.), Computerized adaptive testing: A primer (pp. 101-133). Mahwah, NH: Lawrence Erlbaum Associates, Inc.

Thompson, N. A. (2007). A practitioner's guide for variable-length computerized classification testing. *Practical Assessment Research & Evaluation, 12*(1).

Thompson, N. A., & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation*, *16*.

van der Linden, W. J. (1998). Bayesian item-selection criteria for adaptive testing. *Psychometrika*, *62*, 201–216. https://doi.org/10.1007/BF02294775

van der Linden, W. J., & Pashley, P. J. (2009). Item selection and ability estimation in adaptive testing. In *Elements of adaptive testing* (pp. 3-30). Springer, New York, NY. https://doi.org/10.1007/978-0-387-85461-8_1

van der Linden, W. J., & Ren, H. (2015). Optimal Bayesian adaptive design for test-item calibration. *Psychometrika, 80*(2), 263-288. https://doi.org/10.1007/s11336-013-9391-8

Veerkamp, W. J. J. & Berger, M. P. F. (1997). Item-selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics*, *22*, 203–226. https://doi.org/10.3102/10769986022002203

Verschoor, A., Berger, S., Moser, U., & Kleintjes, F. (2019). On-the-Fly Calibration in Computerized Adaptive Testing. In *Theoretical and Practical Advances in Computer-based Educational Measurement* (pp. 307-323). Springer, Cham. https://doi.org/10.1007/978-3-030-18480-3_16

Vispoel, W. P., Rocklin, T. R., & Wang, T. (1994). Individual differences and test administration procedures: A comparison of fixed-item, computerized-adaptive, and self-adapted testing. *Applied Measurement in Education, 7*(1), 53-79. https://doi.org/10.1207/s15324818ame0701_5

Wainer, H. (1993), Some Practical Considerations When Converting a Linearly Administered Test to an Adaptive Format. *Educational Measurement: Issues and Practice, 12*: 15-20. https://doi.org/10.1111/j.1745-3992.1993.tb00519.x

Wainer, H., & Dorans, N. J. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Mahwah: Lawrence Erlbaum Associates. https://doi.org/10.4324/9781410605931

Wainer, H., & Eignor, D. (2000). Caveats, pitfalls, and unexpected consequences of implementing largescale computerized testing. In H. Wainer, Dorans, N. J., R. Flaugher, B. F. Green, & R. J. Mislevy (Eds.), *Computerized adaptive testing: A primer* (pp. 171-199). Mahwah, NJ: Erlbaum. https://doi.org/10.4324/9781410605931

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. *Computerized adaptive testing: A primer*, *4*, 65-102.

Wang, C., & Xu, G. (2015). A mixture hierarchical model for response times and response accuracy. *British Journal of Mathematical and Statistical Psychology, 68,* 456-477. https://doi.org/10.1111/bmsp.12054

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). Improving measurement quality and efficiency with adaptive testing. *Applied psychological measurement, 6*(4), 473-492. https://doi.org/10.1177/014662168200600408

Weiss, D. J., & Guyer, R. (2012). *Manual for CATSim: Comprehensive simulation of computerized adaptive testing.* St. Paul, MN: Assessment Systems Corporation.

Weiss, D. J., & Kingsbury, G. G. (1984). Application of computerized adaptive testing to educational problems. *Journal of Educational Measurement*, *21*(4), 361-375. https://doi.org/10.1111/j.1745-3984.1984.tb01040.x

Zheng, Y. (2014). *New methods of online calibration for item bank replenishment.* (Doctoral dissertation). University of Illinois, Urbana-Champaign.

Zheng, Y. (2016). Online calibration of polytomous items under the generalized partial credit model. *Applied Psychological Measurement, 40*(6), 434-450. https://doi.org/10.1177/0146621616650406

Zheng, Y., & Chang, H. H. (2017). A comparison of five methods for pretest item selection in online calibration. *International Journal of Quantitative Research in Education*, *4*(1-2), 133-158. https://doi.org/10.1504/IJQRE.2017.086500

Zhu, R. (2006). *Implementation of optimal design for item calibration in computerized adaptive testing (cat).* (Doctoral dissertation). University of Illinois, Urbana-Champaign.

Zimowski, M. F., Muraki, E., Mislevy, R. J. & Bock, R. D. (1996). *Bilog MG: Multiple-group IRT analysis and test maintenance for binary items.* Chicago: Scientific Software International, Inc.

APPENDIX-A: Example Codes of the Online Calibration Program

Example Code for IRT functions (Item Probality Function)

```
NumericVector Prob0_mRcpp(double theta, NumericMatrix items, double D=1) {
  int count = items.nrow();
  NumericVector aux0(count);
  NumericVector pr(count);


  for(int i = 0; i < count; ++i)
  {
    aux0[i] = exp(D*items(i,0)*(theta-items(i,1)));
    pr[i] = (items(i,2) + (1-items(i,2))*(aux0[i]/(1+aux0[i])));
  }
  return pr;
}
```

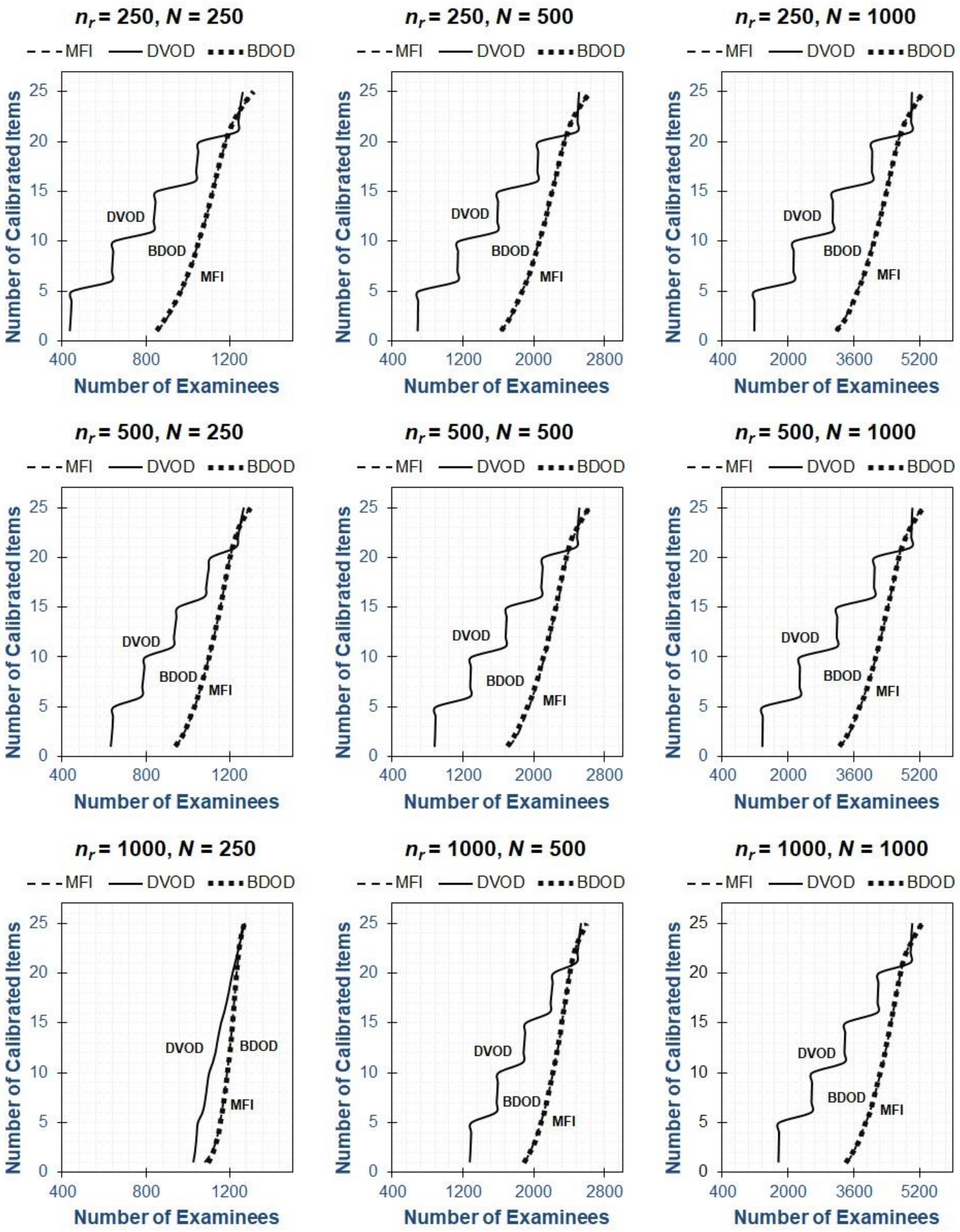
Example Code for CAT functions (Select First Item/s)

```
List FirstItemSelect_mRcpp(NumericMatrix it_pool, NumericVector theta =
NumericVector::create(0) , double D = 1, String rule = "RNG", int n_rand = 1,
Rcpp::Nullable<double> seed = R_NilValue, NumericVector range_it =
NumericVector::create(-0.5,0.5)){

  if (rule != "DFC" && rule != "MFI" && rule != "RNG" )
  {
    stop("Check start item selection rule Diffuculty 'DFC', Maximum Fisher
Information 'MFI' or Range 'RNG'\n");
  }

  if(rule == "RNG" && (range_it.length()!=2 || range_it.isNULL() || range_it[1]<
range_it[0]))
  {
    stop("If rule 'RNG', define 'range_it' correctly!\n");
  }

  NumericVector fit;
  double nit;
  nit = it_pool.nrow();
  double l_theta;
  l_theta = theta.length();
  NumericMatrix fit_par(l_theta,3);

          ▪          ▪          ▪

          ▪          ▪          ▪

          ▪          ▪          ▪

   List firstitem;

   if(sum(!is_na(fit))==0)
   {
     firstitem=List::create(Named("fit")=R_NilValue, Named("fit_par")=R_NilValue,
Named("theta")=theta, Named("rule")=rule);
   }
   else
   {
     firstitem=List::create(Named("fit")=fit[!is_na(fit)],
Named("fit_par")=Valid_mRcpp(fit_par, fit) , Named("theta")=theta,
Named("rule")=rule);
   }

    return firstitem;
}
```

Example Code for Online Calibration functions (Compute D-Optimal Value)

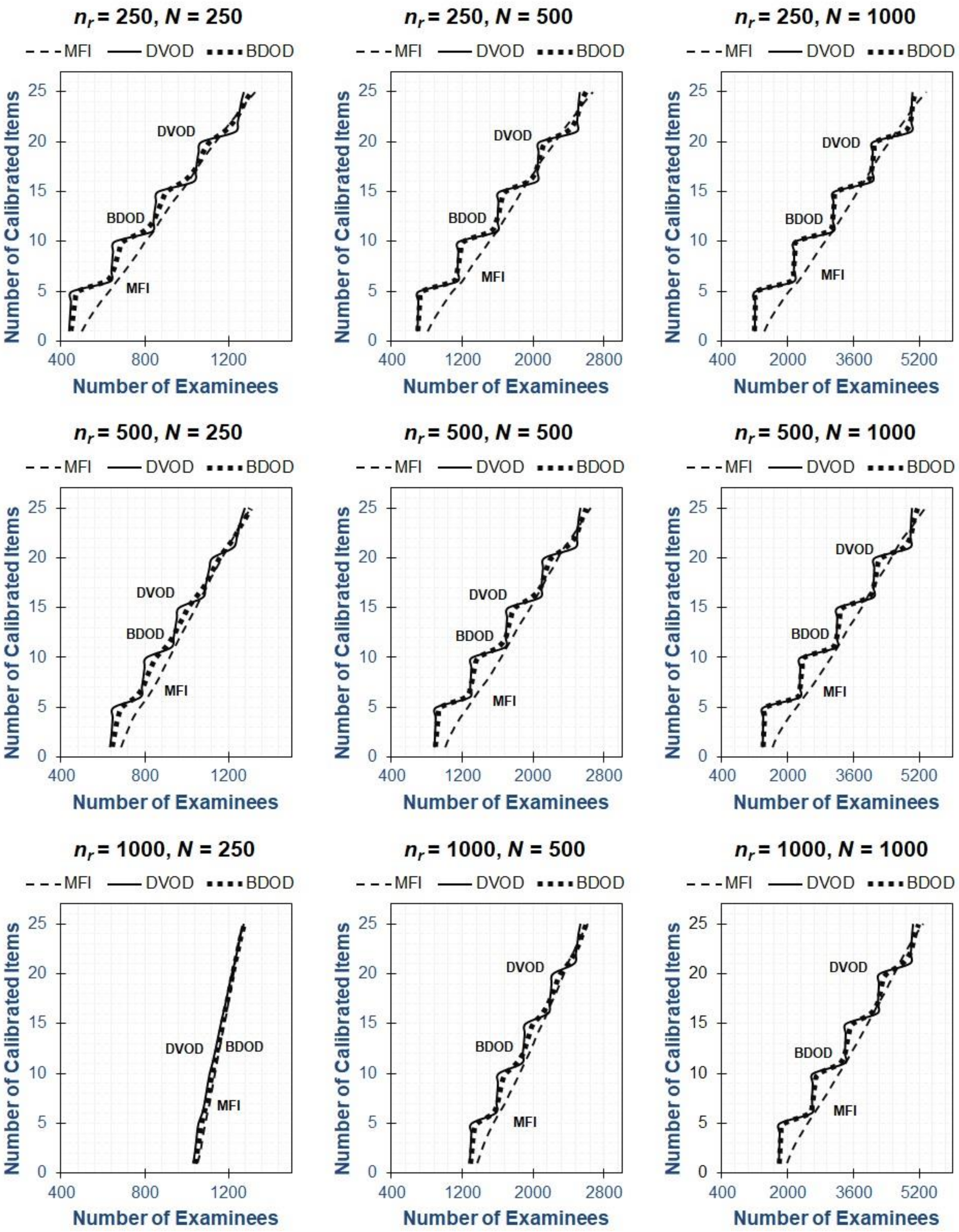
```
double DOptV_mRcpp(NumericVector adm_theta, NumericVector it_par, double
D=1, String irt_model="2PL")
{
  if (irt_model != "1PL" && irt_model != "2PL")
  {
    stop("Check IRT model 'irt_model'\n");
  }

  double dopt;
  dopt=0;
  int l_alltheta=adm_theta.length();
  NumericMatrix item_mpar;
  item_mpar=VecM_mRcpp(it_par);

  if(irt_model=="1PL")
  {
    arma::mat dmat(1,1);
    dmat.zeros();
    for(int i=0;i<l_alltheta; i++)
    {
      NumericVector P;
      NumericVector Q;
          .       .       .
          .       .       .
          .       .       .
    }
    dopt=arma::det(dmat);
  }
  else(irt_model=="2PL")
  {
    arma::mat dmat(2,2);
    dmat.zeros();
    for(int i=0;i<l_alltheta; i++)
    {
      NumericVector P;
      NumericVector Q;
          .       .       .
          .       .       .
          .       .       .
    }
    dopt=arma::det(dmat);
  }

  return dopt;
}
```

APPENDIX B: Figures of The Cumulative Sample Size of Pretest Items for JML Methods Under 1-PL Model



Note. The lines of MFI and BDOD overlap.

APPENDIX C: Figures of The Cumulative Sample Size of Pretest Items for JML Methods Under 2-PL Model


n_r = 250, N = 250
n_r = 250, N = 500
n_r = 250, N = 1000
n_r = 500, N = 250
n_r = 500, N = 500
n_r = 500, N = 1000
n_r = 1000, N = 250
n_r = 1000, N = 500
n_r = 1000, N = 1000
MFI
DVOD
BDOD
Number of Calibrated Items
Number of Examinees
DVOD
BDOD
MFI