International Journal of Assessment Tools in Education
2022, Vol. 9, No. 3, 654–681
https://doi.org/10.21449/ijate.1070848
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Research Article*

The study of the effect of item parameter drift on ability estimation obtained from adaptive testing under different conditions

Merve Sahin Kursad[1,*], **Omay Cokluk Bokeoglu**[2], **Rahime Nukhet Cikrikci**[2]

[1]National Defense University, Department of Measurement and Evaluation, Ankara, Türkiye
[2]Ankara University, Faculty of Education, Department of Educational Sciences, Ankara, Türkiye
[3]İstanbul Aydın University, Faculty of Science and Literature, İstanbul, Türkiye

ARTICLE HISTORY

Received: Feb. 09, 2022
Revised: July 25, 2022
Accepted: Aug. 08, 2022

Keywords:
Item parameter drift, Computer adaptive test, Measurement precision, Test information function.

Abstract: Item parameter drift (IPD) is the systematic differentiation of parameter values of items over time due to various reasons. If it occurs in computer adaptive tests (CAT), it causes errors in the estimation of item and ability parameters. Identification of the underlying conditions of this situation in CAT is important for estimating item and ability parameters with minimum error. This study examines the measurement precision of IPD and its impacts on the test information function (TIF) in CAT administrations. This simulation study compares sample size (1000, 5000), IPD size (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), percentage of items containing IPD (0%, 5%, 10%, 20%), three time points and item bank size (200, 500, 1000) conditions. To examine the impacts of the conditions on ability estimations; measurement precision, and TIF values were calculated, and factorial analysis of variance (ANOVA) for independent samples was carried out to examine whether there were any differences between estimations in terms of these factors. The study found that an increase in the number of measurements using item bank with IPD items results in a decrease in measurement precision and the amount of information the test provides. Factorial ANOVA for independent samples revealed that measurements precision and TIF differences are mostly statistically significant. Although all IPD conditions negatively affect measurement precision and TIF, it has been shown that sample size and item bank size generally do not have an increasing or decreasing effect on these factors.

1. INTRODUCTION

Computer adaptive tests (CAT) produce more reliable results in ability estimations of individuals compared to paper-and-pencil tests and have many advantages. CAT administrations based on Item Response Theory (IRT) place each individual's ability on the same scale with item difficulty values by employing a variety of computer algorithms and measuring the probability that 50% of individuals will provide a correct response to the relevant item (Lord, 1980; Reckase, 2011). This way, tests can be conducted that are more efficient than paper-and-pencil tests in terms of cost and time, but are just as valid and reliable as paper-and-pencil tests by providing individuals with suitable items in line with their ability levels (Çıkrıkçı-Demirtaşlı, 1999; Kaptan, 1993; Wainer, 1993; Weiss & Kingsbury, 1984).

*CONTACT: Merve ŞAHİN KÜRŞAD ✉ sahinmerv@gmail.com Devlet Mahallesi, Kara Harp Okulu Caddesi, National Defense University, Department of Measurement and Evaluation, Ankara, Türkiye

e-ISSN: 2148-7456 /© **IJATE** 2022

Creating a large item bank consisting of high-quality items in CAT administrations is the primary step and an important factor for obtaining valid and reliable results. During the administration of tests, it is important for these items to be of high quality and to maintain this characteristic in successive administrations to obtain accurate results (Bock et al, 1988). Maintaining the item bank's continuity is important for test reliability and observing the changes in item parameters (Risk, 2015). The long-term use of items in the item bank may negatively affect the quality of items, because the repeated use of certain items in administration results in individuals becoming familiar to with these items. Even if the reliability of the item bank is ensured, the frequent encounter of individuals with the same items becomes a factor that compromises reliability and causes item parameters to change or deviate from their original values over time. This change is called item parameter drift (IPD) (Bock et al., 1988). First introduced to the literature in the 1980s, IPD is defined as the differentiation of item parameters over time in successive administrations of tests (Hatfield & Nhouyvanisvong, 2005; McCoy, 2009). This differentiation may occur in one or more parameters of an item (Goldstein, 1983).

Item parameter drift may even occur in situations where the security of the item bank is ensured, and high-quality items are prepared. There are several reasons for the occurrence of IPD in items. Some of these reasons may be listed as: historical and cultural changes, incorrect item calibration, miscalculation of item location on the scale, changes in knowledge, skills, and educational activities, overuse of items, changes in policy or curricula, cheating or security (Li, 2008; Stahl & Muckle, 2007). IPD arising from these reasons may increase or decrease item difficulty or simultaneously increase and decrease item difficulty or other item parameters. Certain negative results thus may arise. The most significant of these negative results is the violation of the invariance assumption, one of the basic assumptions of IRT. If the invariance assumption is ensured, the differences between scores accurately reflect ability differences between individuals or individuals' development over time. However, the occurrence of IPD leads to errors in measurement results, and the test may measure something outside of the construct it intends to measure. Validity also decreases when variables that are irrelevant to the measured construct get mixed into measurement results (McCoy, 2009). This leads to certain problems in the administrations of tests that require the invariance property in item parameters, including test equating, test developing/parallel test developing and CAT (Li, 2008). For instance, in the event of IPD in pre-test items of CAT administrations, errors may occur in item calibration (Meng et al., 2010).

When the scores of two individuals are close to each other, or an individual's score is close to the cut-off score, IPD may lead to incorrect pass-fail decisions and deviations in ability estimations (Rupp & Zumbo, 2006). Apart from that, IPD can also occur when there is not enough time to answer the questions in a test. Not providing sufficient time results in individuals being unable to reach certain items at the end of the test and these items appear more difficult than they actually are. If this problem persists in successive tests administrations, errors pile up and the measurement using previous test items is negatively affected. This leads to the measurement scale to drift (Wise & Kingsbury, 2000).

Another impact of IPD can be observed in CAT administrations. Similar to paper-and-pencil tests, both item and ability parameters are negatively affected in CAT administrations. In terms of item parameters, using previous item parameter estimations to scale new test items leads to errors in item parameter calibration. This results in the deviation of item parameters. The deviation of items from their original parameter values results in the incorrect calibration of pre-test items, leading to errors in individuals' ability estimations (Deng & Melican, 2010). As CAT administrations become more frequently used, the occurrence of IPD in these administrations negatively affects the accuracy of ability estimations and the validity of inferences from test scores.

IPD is a condition that affects the accuracy of individuals' ability estimations and pass/fail decisions. Examining the effect of IPD on measurement precision and TIF is crucial for the safe combination of exams as a whole and the validity of inferences to be made from test scores. Although the use of CAT is widespread, the presence of IPD in these applications negatively affects the accuracy of ability estimations and the validity of inferences made from test scores. Therefore, an examination of the impact of IPD on ability estimations for CAT administrations is significant for the validity of inferences from test scores. Additionally, items containing IPD may have different effects in groups participating in different test administrations. This is a significant issue for CAT administrations since it violates the invariance assumption, one of the basic assumptions of IRT (Babcock & Albano, 2012). The presence of IPD in test applications, where large item banks are used, and especially important decisions are made about the test takers, causes variables unrelated to the structure to interfere with the measurement results, thus reducing the validity. This issue negatively affects measurement precision of scores and validity when interpreting scores in particular (Risk, 2015). For this reason, carrying out IPD studies of item banks in CAT administrations serves to counter this issue, posing threats to construct validity (Wainer et al., 2010). The results of this study are also important to see how the direction, amount and size of the deviations in the item difficulty parameter affect measurement precision and TIF for future CAT applications. In this direction, it is expected that the research findings will provide psychometric information about the organization of the CAT, the sustainability, and updating of the item bank to the institutions and organizations serving in the field of measurement and evaluation.

The overuse of items in successive CAT administrations is a significant cause of IPD occurrence (Bock et al., 1988). For this reason, the item bank should regularly be inspected and updated. IPD studies should therefore be conducted for CAT administrations. However, few studies in the literature examine the impacts of IPD on estimation of ability and item parameters in CAT administrations (Aksu Dünya, 2017; Deng & Melican, 2010; Guo & Wang, 2003; Han & Guo, 2011; Risk, 2015). When some of these studies were closely examined, Guo and Wang (2003) examined the effect of scale drift on the CAT application. The study was conducted with real and simulative data, and the bias in ability estimations and the change in test scores were calculated. Bias, test characteristic curves, and item characteristic curves were compared. As a result of the research, it was stated that a low amount of bias was observed, and this was not important in practical terms. In addition, it was determined that scale drift affects test scores, but this change between two time points is very low. Deng and Melican (2010) studied IPD at multiple time points in CAT applications. The adaptive ACCUPLACER® test was evaluated as part of the scope of the study. Four time points were analyzed using a 3-parameter logistic model (3PLM), and the IPD at parameters a, b, and c was examined. In the evaluation, the item and test characteristic curves were compared. As a result of study, very few items were found to have IPD, but none of the items showed IPD due to its frequent occurrence.

Han and Guo (2011) studied IPD in the context of CAT, resulting from practice and curriculum change. In the study, the effect of IPD on item calibration and ability estimation was examined, using both real and simulative data. Items were calibrated according to 3PLM. According to the results of the study, it was determined that the effect of IPD on item calibration and ability estimations was high, but this effect was not statistically significant. A similar result was obtained by Risk (2015) who examined the effect of IPD on ability estimations in CAT application under various simulative conditions. The Rasch model was used in the study, and the effect of IPD on measurement precision and test effectiveness was examined. When the findings obtained from all conditions are evaluated in general, it is concluded that there are negligible differences between the baseline data set and the conditions that create IPD. However, the most important finding that emerged as a result of the study was; that IPD size has a greater effect on measurement precision than the number of items showing IPD.

Aksu Dünya (2017) investigated the effect of IPD on ability estimations and classification accuracy in the CAT under the condition that IPD affects subgroups with Rasch dichotomous model. According to the study's findings, classification accuracy was significantly affected when a certain group of individuals were exposed to items with IPD. At the same time, average ability estimates were less affected by IPD. In summary, these studies generally focus on the impacts of IPD on item and ability parameters. While some studies find that IPD has a significant effect on CAT-obtained ability estimates (Abad et al., 2010; Hagge et al., 2011; Risk, 2015), others argue that its effect on CAT-obtained ability estimates is small and insignificant (Aksu Dünya, 2017; Deng & Melican, 2010; Guo & Wang, 2003; Han & Guo, 2011; Jiang et al., 2009; McCoy, 2009). These studies mostly examine the impacts of IPD for two time points. However, to be able to observe the impacts of IPD, measurements should be taken for more than two time points. Because it is stated in the literature that if there is an IPD, its effect can be observed clearly after two time points, the IPD's effect can be observed after two time points. Therefore, more than two time points are needed (Babcock & Albano, 2012; Chan et al, 1999; Deng & Melican, 2010; Kim & Cohen, 1992). During the literature review, we could not find any study examining the impact of this issue on ability estimations and test information function while accounting for sample size, item bank size, and various conditions of IPD. Therefore, the impacts of IPD on factors as mentioned above in CAT administrations are not fully known. The aim of this study is to investigate the impact of IPD on measurement precision and TIF in CAT administrations. To this end, answers to the following research questions are sought:

1. When the sample size is 1000, IPD size is 0.00, 0.50, 0.75, 1.00 logit, percentage of items containing IPD is 0%, 5%, 10%, 20%, item bank size is 200, 500, 1000, and measurements are taken for three time points, how do the values of measurement precision and TIF vary in CAT administrations?

2. When the sample size is 5000, IPD size is 0.00, 0.50, 0.75, 1.00 logit, percentage of items containing IPD is 0%, 5%, 10%, 20%, item bank size is 200, 500, 1000, and measurements are taken for three time points, how do the values of measurement precision and TIF vary in CAT administrations?

2. METHOD

2.1. Research Model

This is a simulation-based study that utilized simulated data. Simulation studies are frequently favored in real-world situations involving relatively complex processes, implementation issues, or when real data suited to the type of problem are unavailable. Simulation studies consist of data generating and analysis processes appropriate to situations encountered in real life (Burton et al., 2006; Ranganathan & Foster, 2003). Simulated data are frequently preferred, given the fact that most CAT administrations have implementation problems and require a large sample size and a large item bank (Barrada et al., 2010; Kalender, 2011; McDonald, 2002; Patton et al., 2013; Scullard, 2007; Wang et al., 2012). In this study, because small, medium, and especially large item pools and small and especially large sample sizes are used and drawing on IRT, examines certain IPD situations under controlled conditions in CAT administrations, it is a simulative research.

2.2. Data Generation and Analysis

This study used the R programming language and carried out analyses by generating data using the R Studio 3.3.2 CRAN package (Nydick, 2015). The characteristics of CAT administrations and large-scale assessments were considered when generating data. IPD size and the percentage of items containing IPD were considered when creating conditions for IPD. Also, data was created for taking measurements at three time points. The initial data set that does not contain

IPD was used as the baseline data set during data generation, and data sets containing IPD were compared to this baseline data set. Table 1 displays the controlled and manipulated conditions used in data generation.

Table 1. *Controlled and manipulated conditions in simulated data generation.*

Controlled Conditions	Manipulated Conditions
1. Distribution of ability parameters 2. IRT model and distribution of item parameters 3. Direction and type of IPD 4. CAT Conditions • Method of ability estimation • Starting Rule • Method of item selection • Termination Rule	1. Sample size (1000, 5000) 2. IPD size (0.00, 0.50, 0.75, 1.00) 3. Percentage of items containing IPD (0%, 5%, 10%, 20%) 4. Three time points 5. Item bank size (200, 500, 1000)

2.3. Controlled Conditions in Simulated Data Generation

Since the CAT administration in this study used the Bayesian Expected A Posteriori (EAP) estimate for ability estimation, the distribution of ability parameters was generated with normal distribution with a mean of zero and standard deviation of one. Rasch was chosen as the IRT model because it is favored in large-scale assessments such as Trends in International Mathematics and Science Study (TIMMS) and Programme for International Student Assessment (PISA) (Schulz & Frallion, 2009) and IPD studies for large-scale assessments (Babcock & Albano, 2012; Bergstrom et al., 2001; Hagge et al., 2011; Jones & Smith, 2006; Kingsbury & Wise, 2011; McCoy, 2009; Meyers et al., 2009; Witt et al., 2003).Taking into account the characteristics of CAT administrations and studies in the relevant literature (Filho et al., 2014; Svetina et al., 2013), the distribution of item difficulty parameters was generated with normal distribution with a mean of zero and a standard deviation of one.

The study examines the impact of the item difficulty parameter drift towards the easier condition on ability estimations. There are several reasons for examining this condition. These reasons may be listed as: this situation being encountered more frequently (Babcock & Albano, 2012; Hagge et al., 2011; Risk, 2015; Stahl & Muckle, 2007) and situations with drift in item difficulty parameter being more significant than other parameters (Bock et al., 1988; Donoghue & Isham, 1998; Song & Arce-Ferrer, 2009). Another reason is that although frequent exposure to items or factors such as cheating are observed more frequently, these situations negatively affect ability estimations by causing deviation towards the easier (Risk, 2015; Wells et al., 2012).

After IPD conditions were prepared, conditions for CAT administration were formed. Expected A Posteriori (EAP) was used as the ability estimation method. The ability estimation methods frequently used in CAT applications are the Maximum Likelihood Estimation (MLE), EAP and Maximum A Posteriori (MAP) methods. In most of the studies, the EAP ability estimation method yielded better results than the other two methods (Eroğlu, 2013; Kezer, 2013; Keller, 2000; Kingsbury & Zara, 2009; Wang et al., 2012), with a lower standard error (Wang, 1997) and lower bias value than the MLE method (Eroğlu, 2013). The MLE method was not specified as effective because it estimates ability with more items than EAP and MAP methods (Kezer, 2013). For these reasons, the EAP method was used as an ability estimation method in the CAT application.

Prior θ distributions according to scores individuals acquired in pre-tests were used as starting rule. When the Bayesian approach is used as an ability estimation method, the initial θ level is

estimated from the pre-test before estimating the individuals' real abilities. Thus, the first item to be applied will be the item that gives the most information at the initial θ level (Eroğlu, 2013; Kezer, 2013; Segall, 2004). Accordingly, in this study, as the ability estimation method, one of the Bayes methods, EAP, was used, and the prior θ distributions were used as the starting rule according to the scores of the individuals from the pre-test. The Kullback–Leibler divergence was used for the item selection method. Basic item selection methods used in CAT applications; Maximum Fisher Information, Kullbak-Leibler Information, Interval Information Criterion, Likelihood Weighted Information Criterion, a-stratification, Gradual Maximum Information Ratio, Optimal -b Value (Sulak, 2013). In studies comparing the performance of these methods, -a stratification and Kullbak-Leibler item selection methods have better performances in ability estimations than other methods (Barrada et al., 2010; Chang & Ying; 1999; Chen et al., 2000; Deng et al., 2010; Eggen, 1999; Linda 1996; Sulak, 2013; Veldkamp & van der Linden, 2006; Yao, 2013). However, since the analyzes were made based on the Rasch model within the scope of this study, the -a stratification method is not suitable because the discrimination values of all items are constant. For this reason, the Kullbak-Leibler item selection method was preferred.

Lastly, the minimum number of items rule, one of the variable-length termination rules, and standard error were used as the termination rule. For the minimum number of items rule, the termination rule was set as minimum 10 items and standard error at less than 0.40. Higher error and bias values are obtained when the minimum number of items applied is less than 10 (Babcock & Weiss 2012; Erolu, 2013), and the normal distribution is compromised when the minimum number of items applied is low (Blais & Raiche, 2002). Therefore, in this study, a minimum of 10 items was preferred for the minimum number of items rule. In the standard error termination rule between the [-3.00; +3.00] ability interval, a standard error equal to or less than 0.40 is suitable for measurement precision (Babcock & Weiss, 2012; Blaise & Raiche, 2002). Therefore, these termination rules were preferred.

2.4. Manipulated Conditions in Simulated Data Generation

While a sample size of 1000-2000 is required to make accurate estimations of item parameters based on IRT (Rudner & Guo, 2011; Stahl & Muckle, 2007), lower standard error values are obtained when the sample size is 5000 (Şahin, 2012). The sample size of 1000 was thus treated as the small sample size and 5000 as the large sample size. One of the most important factors affecting the estimation of ability is the size of the IPD (Risk, 2015). IPD size of 0.50 logit or more significantly affects parameter estimations (Donoghue & Isham, 1998; Han & Wells, 2007; Wollack et al., 2005). Therefore 0.00, 0.50, 0.75, and 1.00 logit were generated as IPD magnitude to examine the impact of IPD magnitude.

As one of the factors negatively affecting ability estimations, the IPD percentage (Hagge et al., 2011; Huang & Shyu, 2003; Wells et al., 2002) was found to range between 5 and 20–25% in the relevant literature (Hagge et al., 2011; Stahl et al., 2002; Song & ArceFerrer, 2009; Wells et al., 2002). This study examines IPD-containing items with 0%, 5%, 10%, and 20%. To fully reveal the impact of IPD, more than two time points or measurements are needed (Babcock & Albano, 2012; Chan et al., 1999; Deng & Melican, 2010; Kim & Cohen, 1992). For this reason, this study uses parameter estimations at three time points. In line with some studies in the literature regarding item bank size in the CAT application (Han & Guo, 2011; Risk, 2015; Veldkamp & Linden, 2006; Wise & Kingsbury, 2000), this study set item bank sizes of 200, 500 and 1000 for small, medium and large item banks respectively.

Given the controlled and manipulated conditions, simulated data were generated for 288 situations, calculated as 2 (sample sizes) × 3 (item bank sizes) × 4 (IPD sizes) × 4 (IPD percentages) × 3 (time points). For every situation, a total of 100 replications were carried out and 28,800 analyses were performed. In simulation studies, replication numbers must be kept higher to see the effect of the variables on the situations to be observed more clearly (Köse &

Başaran, 2021). As Evans (2010) quoted, to eliminate bias caused by sample size, at least 25 replications were recommended (Harwell, 1996). Consequently, 100 replications were favored. To examine the effect of condition on estimations of ability, values for measurement precision (bias and root-mean-square error -RMSE-) and TIF were calculated. The calculation formulas are displayed in Table 2 below.

Table 2. *Assessment criteria for item parameter drift.*

Criteria		Description	Formula
Measurement Precision	Bias	Systematic deviation of real ability from estimated ability.	$\frac{\sum_{j=1}^{n}(\widehat{\Theta i}-\Theta i)}{n}$
	RMSE	Root mean square error	$\sqrt{\frac{\sum_{j=1}^{n}(\widehat{\Theta i}-\Theta i)^2}{n}}$
Test Information Function	TIF	The test information function is equal to the total information function of items individuals obtain from the relevant test. This value is calculated using standard error values.	$S_{em}(\hat{\theta})=\frac{1}{\sqrt{I(\theta)}}$

θi: Ability of individuals, $\widehat{\theta i}$: Estimated ability of individuals, n: Total number of individuals, I(Θ): Item information function. Also, measurement precision and TIF are correlated with each other by SEM with $RMSE^2 = BIAS^2 + S_{em}^2$ formula

After calculating values for measurement precision and TIF for 100 replications using the formula in Table 2, a three-factor analysis of variance (ANOVA) was performed for independent samples to examine whether the obtained values displayed statistically significant differences. In the analysis, the independent variables consisted of IPD size (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), IPD percentage (0%, 5%, 10%, 20%), and measurements using item banks with IPD (3 measurements), while the dependent variables consisted of bias, RMSE, and TIF values. Along with ANOVA, the Eta squared (η^2) effect size was also reported. When interpreting the effect size, .01 was taken as small, .09 as a medium, and .25 as large effect sizes (Cohen, 1988). When calculating the impact of IPD for every condition, the initial data set that did not contain IPD was taken as the baseline data set. After forming IPD conditions using this data set, data sets containing IPD and the baseline data set were compared, and the results were interpreted.

3. FINDINGS

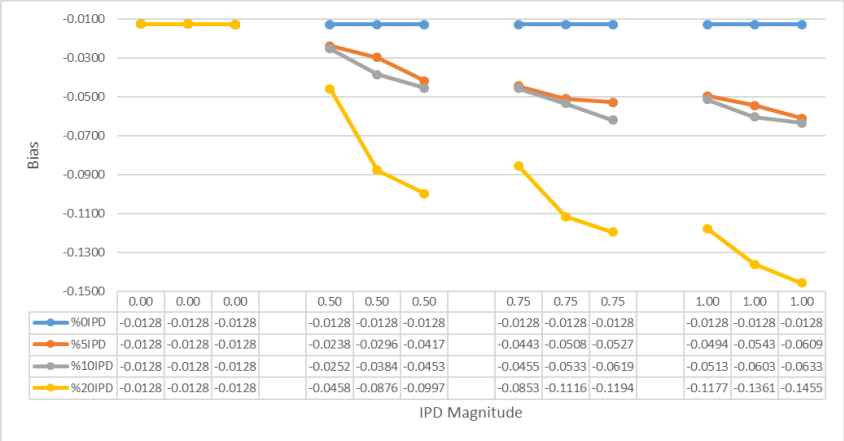
This section first discusses the findings and interpretations obtained from data for the sample size of 1000, then goes on to findings and interpretations of data with a sample size of 5000.

3.1. Findings on Comparison of Conditions with Sample Size 1000

The first criterion for measurement precision, i.e., the dependent variable, is the bias values regarding ability estimations. Findings of bias values are shown in Figure 1. a, b and c. When bias values were examined for a sample size of 1000, the increase in IPD size (0.00, 0.50, 0.75, 1.00) and IPD percentage (0%, 5%, 10%, 20%) for item bank sizes of 200, 500 and 1000, resulted in a tendency of ability estimation *bias* values obtained at three time points to increase in the negative direction. Besides this, as item bank size increased, no increasing or decreasing bias tendency were observed. Negative bias values mean that individuals' estimated ability values are lower than their real ability values. Since certain items in the item bank displayed IPD in the easier direction, we would have expected individuals' estimated ability values to be higher than their real ability values; in other words, bias values should have increased in the positive direction. There may be two reasons for obtaining results in the opposite direction.

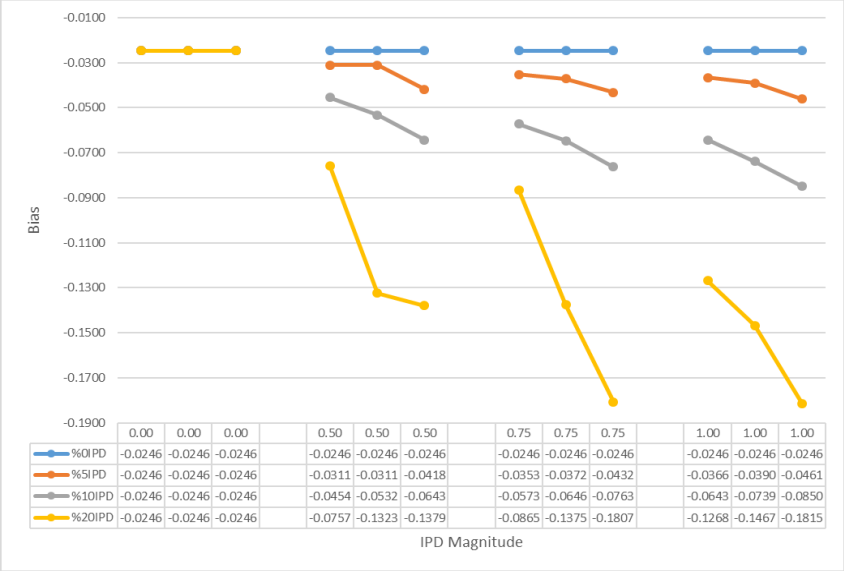
Figure 1. a, b and **c.** *Figures denoting comparison of bias values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=1000.*

a. *Bias values for item bank of 200 with sample size n=1000.*



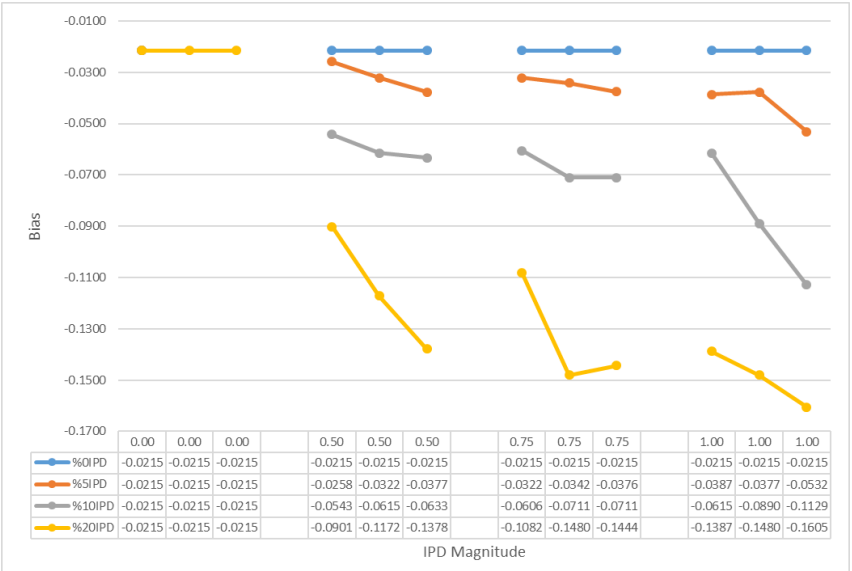
	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128	-0.0128
%5IPD	-0.0128	-0.0128	-0.0128	-0.0238	-0.0296	-0.0417	-0.0443	-0.0508	-0.0527	-0.0494	-0.0543	-0.0609
%10IPD	-0.0128	-0.0128	-0.0128	-0.0252	-0.0384	-0.0453	-0.0455	-0.0533	-0.0619	-0.0513	-0.0603	-0.0633
%20IPD	-0.0128	-0.0128	-0.0128	-0.0458	-0.0876	-0.0997	-0.0853	-0.1116	-0.1194	-0.1177	-0.1361	-0.1455

b. *Bias values for item bank of 500 with sample size n=1000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246	-0.0246
%5IPD	-0.0246	-0.0246	-0.0246	-0.0311	-0.0311	-0.0418	-0.0353	-0.0372	-0.0432	-0.0366	-0.0390	-0.0461
%10IPD	-0.0246	-0.0246	-0.0246	-0.0454	-0.0532	-0.0643	-0.0573	-0.0646	-0.0763	-0.0643	-0.0739	-0.0850
%20IPD	-0.0246	-0.0246	-0.0246	-0.0757	-0.1323	-0.1379	-0.0865	-0.1375	-0.1807	-0.1268	-0.1467	-0.1815

c. *Bias values for item bank of 1000 with sample size n=1000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215	-0.0215
%5IPD	-0.0215	-0.0215	-0.0215	-0.0258	-0.0322	-0.0377	-0.0322	-0.0342	-0.0376	-0.0387	-0.0377	-0.0532
%10IPD	-0.0215	-0.0215	-0.0215	-0.0543	-0.0615	-0.0633	-0.0606	-0.0711	-0.0711	-0.0615	-0.0890	-0.1129
%20IPD	-0.0215	-0.0215	-0.0215	-0.0901	-0.1172	-0.1378	-0.1082	-0.1480	-0.1444	-0.1387	-0.1480	-0.1605

Firstly, IPD in the easier direction occurred only for the item difficulty parameter. If IPD had occurred at both directions, bias estimations would have been calculated as near-zero (Aksu Dünya, 2017; Wei, 2013). Secondly, although individuals were provided with items according to their ability level, they may have answered incorrectly. Some studies in relevant literature have also come up with similar findings (Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003).

On the other hand, a study by Guo and Wang (2003) that examined the impact of the parameter drift in CAT administrations on test scores showed that ability estimation bias values for item banks with IPD were not affected. This is because the study carried out measurements at two time points. Babcock and Albano (2012) also stated that taking ability measurements at two time points is insufficient to make clear inferences about how IPD affects ability estimations. In other words, in order to reveal the effects of IPD, it is necessary to take measurements at least three time points.

Three-factor ANOVA results for independent samples, as shown in Table 3, examine whether obtained differences were statistically significant according to above-mentioned bias values. In Table 3, IPD size represents the drift size of items containing IPD in the item bank (0.00 logit, 0.50 logit, 0.75 logit, 1.00 logit), IPD percentage represents the percentage of items containing IPD in the item bank (0%, 5%, 10%, 20%), and measurement factor represents the number of measurements performed with the item bank containing items with IPD (3 measurements).

Table 3. *Comparison of bias values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=1000.*

Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.515	3	0.172	5366.84*	0.15
	IPD Percentage	2.031	3	0.677	21165.16*	**0.58**
	Measurement	0.235	2	0.118	2448.95*	0.06
	IPD Size*IPD Percentage	0.113	9	0.013	1177.58*	0.03
	IPD Size*Measurement	0.016	6	0.003	166.74*	0.01
	IPD Percentage*Measurement	0.071	6	0.012	739.89*	0.02
	IPD Size*IPD Percentage*Measurement	0.011	18	0.001	114.63*	0.01
	Error	0.456	4752	0.000		
	Total	3.448	4799			
500	IPD Size	0.196	3	0.065	2352.00*	0.03
	IPD Percentage	4.414	3	1.471	52968.00*	**0.73**
	Measurement	0.493	2	0.247	5916.00*	0.08
	IPD Size*IPD Percentage	0.070	9	0.008	840.00*	0.01
	IPD Size*Measurement	0.022	6	0.004	264.00*	0.01
	IPD Percentage*Measurement	0.333	6	0.056	3996.00*	0.05
	IPD Size*IPD Percentage*Measurement	0.060	18	0.003	720.00*	0.01
	Error	0.396	4752	0.000		
	Total	5.984	4799			
1000	IPD Size	0.271	3	0.090	2893.91*	0.05
	IPD Percentage	4.232	3	1.411	45192.05*	**0.78**
	Measurement	0.248	2	0.124	2648.31*	0.04
	IPD Size*IPD Percentage	0.047	9	0.005	501.90*	0.01
	IPD Size*Measurement	0.024	6	0.004	256.29*	0.01
	IPD Percentage*Measurement	0.058	6	0.010	619.36*	0.01
	IPD Size*IPD Percentage*Measurement	0.075	18	0.004	800.90*	0.01
	Error	0.445	4752	0.000		
	Total	5.400	4799			

*$p<.05$

ANOVA results for independent samples regarding bias show that the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects

on bias. These generally have low effect sizes (Cohen, 1988). The post-hoc analysis results also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation bias among the variables within the scope of this study. Aksu Dünya (2017) and Babcock and Albano (2012), who used the Rasch model and Abad et al. (2010), who used the 3PLM IRT model, obtained similar findings.

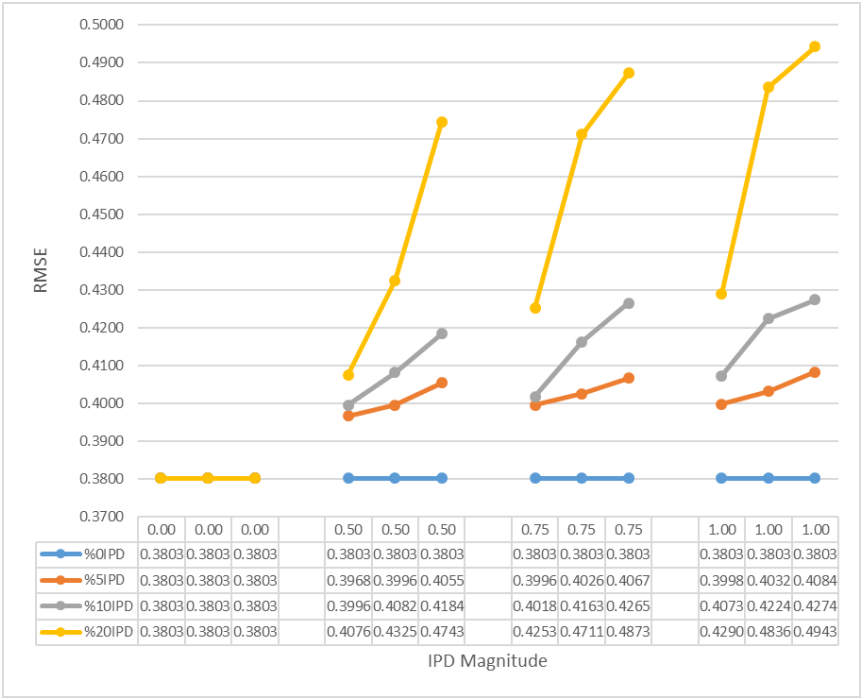
The second criterion for measurement precision, i.e., the dependent variable, is the RMSE values for ability estimations. Obtained RMSE values are shown in Figure 2. a, b and c.

Figure 2. a, b. and **c.** *Figures denoting comparison of RMSE values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=1000.*

a. *RMSE values for item bank of 200 with sample size n=1000.*

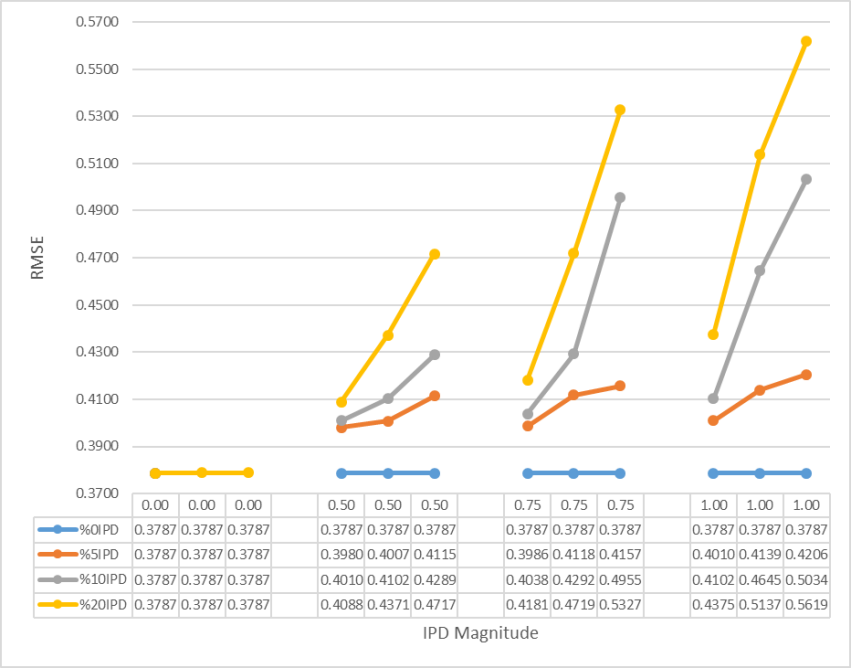
	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958	0.3958
%5IPD	0.3958	0.3958	0.3958	0.3972	0.4049	0.4057	0.4005	0.4046	0.4065	0.4013	0.4062	0.4082
%10IPD	0.3958	0.3958	0.3958	0.3999	0.4079	0.4096	0.4029	0.4106	0.4140	0.4077	0.4121	0.4161
%20IPD	0.3958	0.3958	0.3958	0.4088	0.4408	0.4532	0.4197	0.4515	0.4668	0.4364	0.4530	0.4709

b. *RMSE values for Item Bank of 500 with Sample Size n=1000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803	0.3803
%5IPD	0.3803	0.3803	0.3803	0.3968	0.3996	0.4055	0.3996	0.4026	0.4067	0.3998	0.4032	0.4084
%10IPD	0.3803	0.3803	0.3803	0.3996	0.4082	0.4184	0.4018	0.4163	0.4265	0.4073	0.4224	0.4274
%20IPD	0.3803	0.3803	0.3803	0.4076	0.4325	0.4743	0.4253	0.4711	0.4873	0.4290	0.4836	0.4943

c. *RMSE Values for Item Bank of 1000 with Sample Size n=1000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787	0.3787
%5IPD	0.3787	0.3787	0.3787	0.3980	0.4007	0.4115	0.3986	0.4118	0.4157	0.4010	0.4139	0.4206
%10IPD	0.3787	0.3787	0.3787	0.4010	0.4102	0.4289	0.4038	0.4292	0.4955	0.4102	0.4645	0.5034
%20IPD	0.3787	0.3787	0.3787	0.4088	0.4371	0.4717	0.4181	0.4719	0.5327	0.4375	0.5137	0.5619

When the RMSE values were examined for the sample size of 1000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000, resulted in a tendency of ability estimation RMSE values obtained at three time points to increase. The increase in the number of measurements in item banks containing IPD, IPD size, and IPD percentage results in more erroneous ability estimations leading to a decrease in measurement precision. The lowest values of RMSE were obtained in the baseline data set, since there was no IPD. However, RMSE values decreased as the item bank size increased for the baseline datasets. In other words, as the item bank size increases, less erroneous results regarding ability estimations were obtained in the baseline data set. Besides this, as item bank size increased for data sets with IPD, no increasing or decreasing RMSE tendency were observed. Some studies in relevant literature have also obtained similar findings (Aksu Dünya, 2017; Babcock & Albano, 2012; Chen, 2013; Risk, 2015; Wells et al., 2002). While Aksu Dünya (2017) argues that the lowest RMSE value was obtained for the baseline data set, it is stated that the increase in the percentage of items containing IPD resulted in more erroneous ability estimations. Wells et al. (2012) found that as sample size increased, RMSE values decreased, leading to more accurate estimates. However, as IPD size increased within the same sample size, RMSE values increased, leading to less precise measurements. Three-factor ANOVA results for independent samples are shown in Table 4 which examine whether obtained differences were statistically significant according to the RMSE values discussed above.

Results of a three-factor ANOVA on RMSE values for independent samples indicate that the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on RMSE. These generally possess low and high effect sizes (Cohen, 1988). The results of post-hoc analysis revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation RMSE among the variables within the scope of this study. Risk (2015) also reached similar findings. A study by Babcock and Albano (2012) obtained similar findings, but argued that the factor with the most impact on RMSE values was IPD size.

Table 4. *Comparison of RMSE values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size is n=1000.*

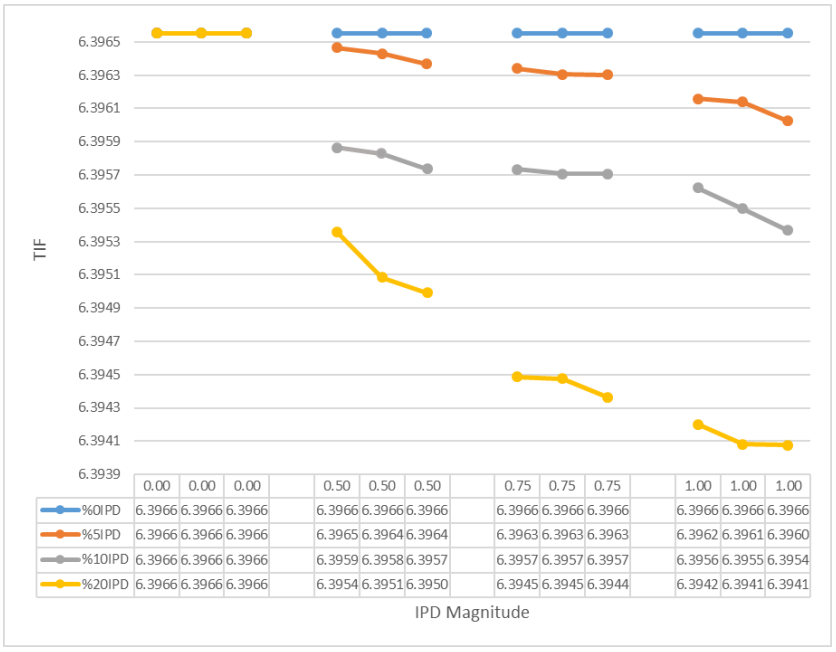
Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η^2)
200	IPD Size	0.039	3	0.013	652.56*	0.02
	IPD Percentage	0.884	3	0.295	14791.44*	**0.57**
	Measurement	0.180	2	0.090	3011.83*	0.11
	IPD Size*IPD Percentage	0.023	9	0.003	384.85*	0.01
	IPD Size*Measurement	0.005	6	0.001	83.66*	0.01
	IPD Percentage*Measurement	0.115	6	0.019	1924.23*	0.07
	IPD Size*IPD Percentage*Measurement	0.005	18	0.000	83.66*	0.00
	Error	0.284	4752	0.000		
	Total	1.535	4799			
500	IPD Size	0.104	3	0.035	1752.51*	0.03
	IPD Percentage	1.431	3	0.477	24113.87*	**0.54**
	Measurement	0.449	2	0.225	7566.13*	0.16
	IPD Size*IPD Percentage	0.068	9	0.008	1145.87*	0.02
	IPD Size*Measurement	0.016	6	0.003	269.62*	0.00
	IPD Percentage*Measurement	0.274	6	0.046	4617.19*	0.10
	IPD Size*IPD Percentage*Measurement	0.020	18	0.001	337.02*	0.00
	Error	0.282	4752	0.000		
	Total	2.644	4799			
1000	IPD Size	0.718	3	0.239	9425.24*	0.12
	IPD Percentage	1.880	3	0.627	24678.90*	**0.32**
	Measurement	1.777	2	0.889	23326.81*	0.30
	IPD Size*IPD Percentage	0.253	9	0.028	3321.15*	0.04
	IPD Size*Measurement	0.205	6	0.034	2691.05*	0.03
	IPD Percentage*Measurement	0.552	6	0.092	7246.14*	0.09
	IPD Size*IPD Percentage*Measurement	0.089	18	0.005	1168.31*	0.01
	Error	0.362	4752	0.000		
	Total	5.836	4799			

*$p<.05$

The third criterion for comparing independent variables discussed in the study is the TIF values. Findings for TIF values are shown in Figure 3. a, b and c. When TIF values were examined for a sample size of 1000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation *TIF* values obtained at three time points to decrease. Therefore, the increase in the number of measurements, IPD size and IPD percentage result in a decrease in the amount of information the test provides. This tendency does not change with an increase in item bank size. Studies in the literature indicate that TIF tend to change even at low levels when IPD is present (Chan et al., 1999; Deng & Melican, 2010; Guo & Wang, 2003).

Figure 3. a, b and **c**. *Figures denoting comparison of TIF values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=1000.*

a. *TIF values for item bank of 200 with sample size n=1000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966	6.3966
%5IPD	6.3966	6.3966	6.3966	6.3965	6.3964	6.3964	6.3963	6.3963	6.3963	6.3962	6.3961	6.3960
%10IPD	6.3966	6.3966	6.3966	6.3959	6.3958	6.3957	6.3957	6.3957	6.3957	6.3956	6.3955	6.3954
%20IPD	6.3966	6.3966	6.3966	6.3954	6.3951	6.3950	6.3945	6.3945	6.3944	6.3942	6.3941	6.3941

b. *TIF values for item bank of 500 with sample size n=1000.*

	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973
%5IPD	6.3973	6.3973	6.3973	6.3973	6.3973	6.3971	6.3970	6.3969	6.3969	6.3969	6.3969	6.3968
%10IPD	6.3973	6.3973	6.3973	6.3968	6.3968	6.3968	6.3967	6.3967	6.3966	6.3965	6.3965	6.3964
%20IPD	6.3973	6.3973	6.3973	6.3962	6.3962	6.3961	6.3959	6.3959	6.3954	6.3946	6.3945	6.3944

c. *TIF values for item bank of 1000 with sample size n=1000.*

	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970	6.3970
%5IPD	6.3970	6.3970	6.3970	6.3967	6.3967	6.3967	6.3966	6.3965	6.3965	6.3964	6.3964	6.3964
%10IPD	6.3970	6.3970	6.3970	6.3962	6.3962	6.3962	6.3960	6.3959	6.3959	6.3957	6.3957	6.3957
%20IPD	6.3970	6.3970	6.3970	6.3956	6.3954	6.3954	6.3954	6.3951	6.3950	6.3948	6.3941	6.3941

The three-factor ANOVA results for independent samples, shown in Table 5, examine whether the differences obtained were statistically significant according to the TIF values discussed above.

Table 5. *Comparison of TIF values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=1000.*

Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η2)
200	IPD Size	19.342	3	6.447	79578.51*	**0.35**
	IPD Percentage	8.060	3	2.687	33161.14*	0.14
	Measurement	1.075	2	0.538	4422.86*	0.02
	IPD Size*IPD Percentage	16.112	9	1.790	66289.37*	0.30
	IPD Size*Measurement	2.148	6	0.358	8837.49*	0.03
	IPD Percentage*Measurement	2.148	6	0.358	8837.49*	0.04
	IPD Size*IPD Percentage*Measurement	4.294	18	0.239	17666.74*	0.08
	Error	1.155	4752	0.000		
	Total	54.334	4799			
500	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.000	316.80*	0.02
	Measurement	0.012	2	0.006	3801.60	-
	IPD Size*IPD Percentage	0.000	9	0.000	0.00*	0.00
	IPD Size*Measurement	0.003	6	0.000	950.40	-
	IPD Percentage*Measurement	0.005	6	0.000	1584.00	-
	IPD Size*IPD Percentage*Measurement	0.008	18	0.000	2534.40	-
	Error	0.015	4752	0.000		
	Total	0.045	4799			
1000	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.001	279.53*	0.00
	Measurement	0.018	2	0.009	5031.53	-
	IPD Size*IPD Percentage	0.068	9	0.007	19008.00*	0.00
	IPD Size*Measurement	0.003	6	0.000	838.59	-
	IPD Percentage*Measurement	0.014	6	0.002	3913.41	-
	IPD Size*IPD Percentage*Measurement	0.008	18	0.000	2236.24	-
	Error	0.017	4752	0.000		
	Total	0.130	4799			

**p<.05*

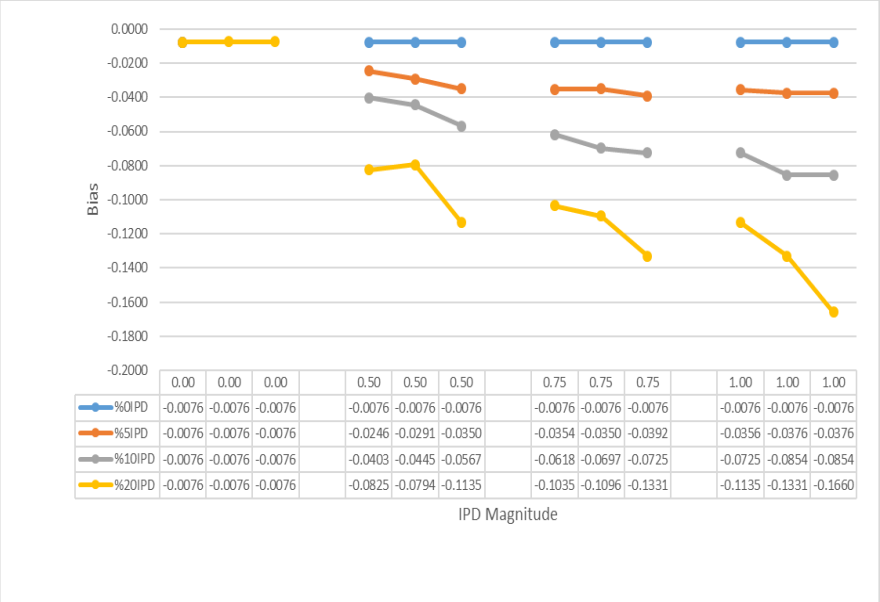
The three-factor ANOVA results for independent samples in terms of TIF values show that the main effect and the effects of the two- and three-factor interactions of the number of measurements, IPD size and IPD percentage have statistically significant effects on TIF for an item bank of 200 items in a sample of 1000. Especially for an item bank of 200 items, IPD size factors significantly affect TIF values. This is high-level effect (Cohen, 1988). Although IPD size, IPD percentage and IPD size*IPD percentage have interaction effects on item banks of 500 and 1000 items, these effects are low-level (Cohen, 1988). While some studies on the impacts of IPD on TIF (Chan et al., 1999) support this finding, some studies argue that there are no statistically significant differences (Deng & Melican, 2010; Guo and Wang, 2003).

3.2. Findings of Comparison of Conditions for Sample Size 5000

The findings for bias values, which constitute the first criterion for comparing independent variable conditions for a sample size of 5000, are shown in Figure 4. a, b and c.

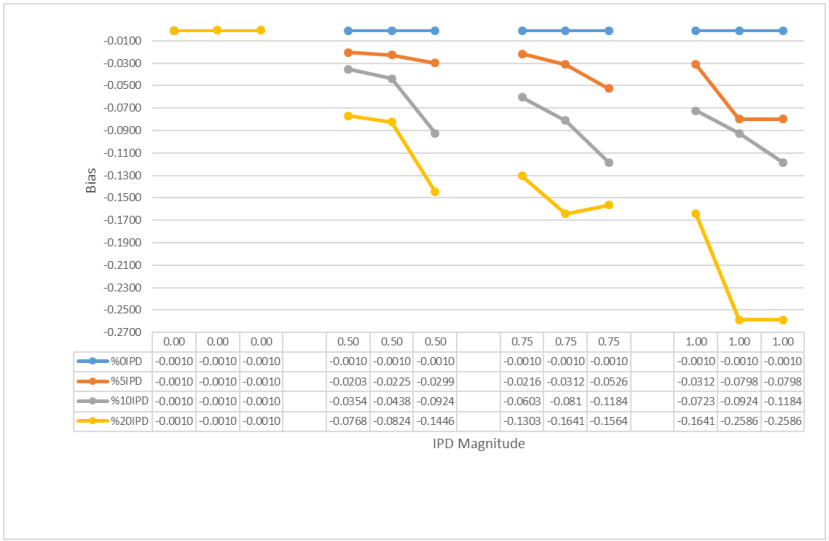
Figures 4. a, b and **c.** *Figures denoting comparison of bias values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=5000.*

a. *Bias values for item bank of 200 with sample size n=5000.*



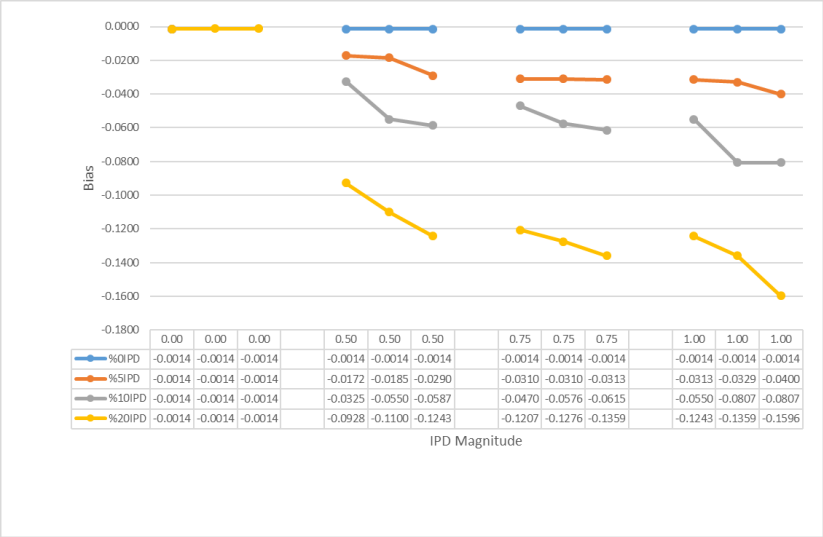
	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076	-0.0076
%5IPD	-0.0076	-0.0076	-0.0076	-0.0246	-0.0291	-0.0350	-0.0354	-0.0350	-0.0392	-0.0356	-0.0376	-0.0376
%10IPD	-0.0076	-0.0076	-0.0076	-0.0403	-0.0445	-0.0567	-0.0618	-0.0697	-0.0725	-0.0725	-0.0854	-0.0854
%20IPD	-0.0076	-0.0076	-0.0076	-0.0825	-0.0794	-0.1135	-0.1035	-0.1096	-0.1331	-0.1135	-0.1331	-0.1660

b. *Bias values for item bank of 500 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010	-0.0010
%5IPD	-0.0010	-0.0010	-0.0010	-0.0203	-0.0225	-0.0299	-0.0216	-0.0312	-0.0526	-0.0312	-0.0798	-0.0798
%10IPD	-0.0010	-0.0010	-0.0010	-0.0354	-0.0438	-0.0924	-0.0603	-0.081	-0.1184	-0.0723	-0.0924	-0.1184
%20IPD	-0.0010	-0.0010	-0.0010	-0.0768	-0.0824	-0.1446	-0.1303	-0.1641	-0.1564	-0.1641	-0.2586	-0.2586

c. *Bias values for item bank of 1000 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014	-0.0014
%5IPD	-0.0014	-0.0014	-0.0014	-0.0172	-0.0185	-0.0290	-0.0310	-0.0310	-0.0313	-0.0313	-0.0329	-0.0400
%10IPD	-0.0014	-0.0014	-0.0014	-0.0325	-0.0550	-0.0587	-0.0470	-0.0576	-0.0615	-0.0550	-0.0807	-0.0807
%20IPD	-0.0014	-0.0014	-0.0014	-0.0928	-0.1100	-0.1243	-0.1207	-0.1276	-0.1359	-0.1243	-0.1359	-0.1596

When bias values were examined for a sample size of 5000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500 and 1000 resulted in a tendency of ability estimation bias values obtained at three time points to grow in the negative direction as in 1000 sample size. Negative bias values mean that individuals' estimated ability values are lower than their real ability values. Since certain items in the item bank displayed IPD in the easier direction, we would have expected that individuals' estimated ability values to be higher than their real ability values. The reason could be either that IPD occurred only in the easier direction and only in the item difficulty parameter (Aksu Dünya, 2017; Wei, 2013), or individuals were provided items according to their ability level and may have answered them incorrectly (Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003). The increase in the number of measurements, IPD size, and IPD percentage results in more biased ability estimations leading to a decrease in measurement precision. Studies in the literature indicate that IPD negatively affects bias values (Aksu Dünya, 2017; Chen, 2013; Risk, 2015; Rupp & Zumbo, 2003). IPD occurrences at and over 0.50 logit in particular significantly affect parameter estimations (Han & Wells, 2007; Wollack et al., 2005). Since this study also examined conditions with IPD at and over 0.50 logit, differences were obtained in bias values, albeit low.

Three-factor ANOVA results for independent samples, shown in Table 6, examine whether obtained differences were statistically significant according to the bias values discussed above. Three-factor ANOVA results for independent samples regarding bias show that both the main effect and effects of two-way and three-way interactions of the number of measurements, IPD size and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on bias. These generally have low effect sizes (Cohen, 1988). The results of post-hoc analysis also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation bias among the variables within the scope of this study. Some studies in the literature have also reached similar findings (Abad et al., 2010; Babcock & Albano, 2012).

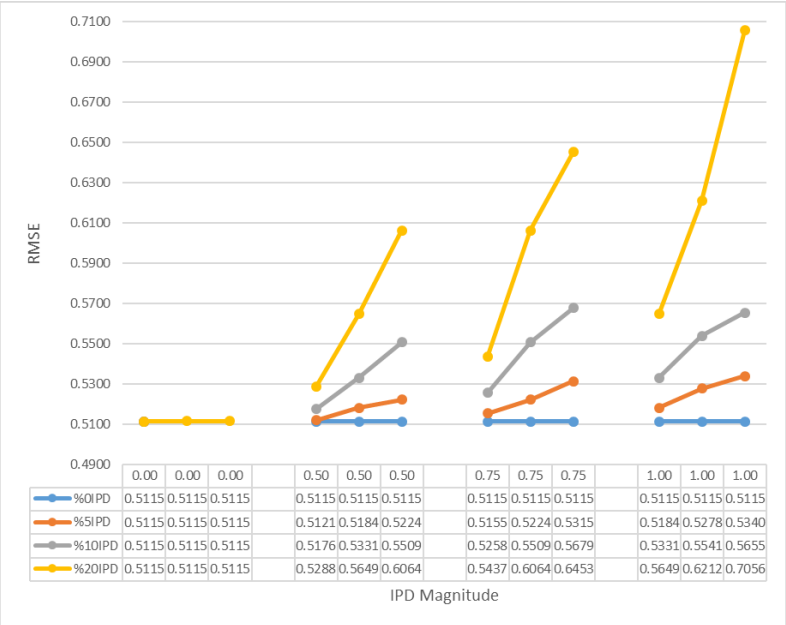
Table 6. *Comparison of bias values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=5000.*

Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η2)
200	IPD Size	0.383	3	0.128	18571.59*	0.10
	IPD Percentage	2.971	3	0.990	144063.18*	**0.76**
	Measurement	0.166	2	0.083	8049.31*	0.04
	IPD Size*IPD Percentage	0.117	9	0.013	5673.31*	0.03
	IPD Size*Measurement	0.011	6	0.002	533.39*	0.00
	IPD Percentage*Measurement	0.104	6	0.017	5042.94*	0.02
	IPD Size*IPD Percentage*Measurement	0.018	18	0.001	872.82*	0.00
	Error	0.098	4752	0.000		
	Total	3.868	4799			
500	IPD Size	2.081	3	0.694	99888.00*	0.18
	IPD Percentage	6.748	3	2.249	323904.00*	**0.60**
	Measurement	1.069	2	0.535	51312.00*	0.09
	IPD Size*IPD Percentage	0.806	9	0.090	38688.00*	0.07
	IPD Size*Measurement	0.224	6	0.037	10752.00*	0.02
	IPD Percentage*Measurement	0.183	6	0.031	8784.00*	0.01
	IPD Size*IPD Percentage*Measurement	0.212	18	0.012	10176.00*	0.02
	Error	0.099	4752	0.000		
	Total	11.422	4799			
1000	IPD Size	0.228	3	0.076	11169.65*	0.04
	IPD Percentage	4.403	3	1.468	215701.61*	**0.88**
	Measurement	0.160	2	0.080	7838.35*	0.03
	IPD Size*IPD Percentage	0.033	9	0.004	1616.66*	0.00
	IPD Size*Measurement	0.017	6	0.003	832.82*	0.00
	IPD Percentage*Measurement	0.049	6	0.008	2400.49*	0.01
	IPD Size*IPD Percentage*Measurement	0.007	18	0.000	342.93*	0.00
	Error	0.097	4752	0.000		
	Total	4.994	4799			

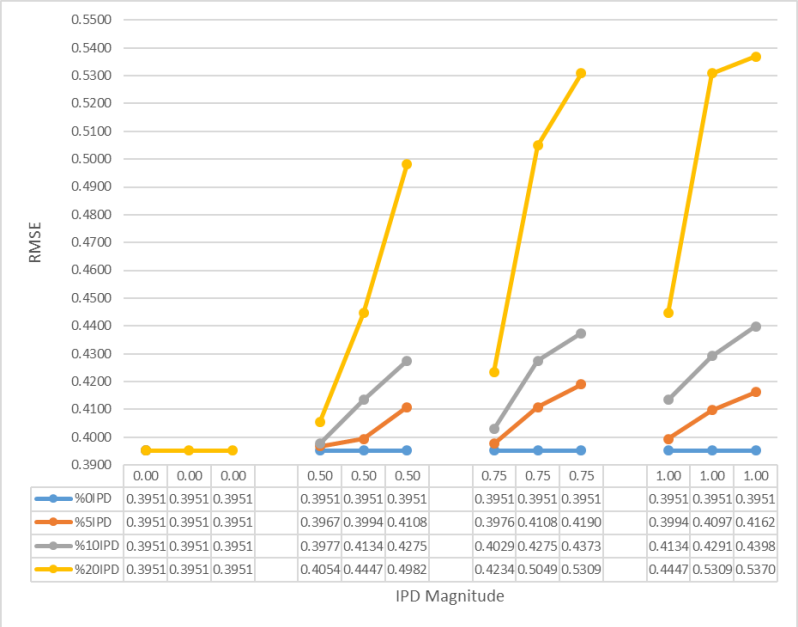
**p<.05*

The findings for RMSE values, which constitute the second criterion for measurement precision where independent variable conditions are compared are shown in Figure 5. a, b and c.

Figure 5. a, b and **c.** *Figures denoting comparison of RMSE values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=5000.*
a. *RMSE values for item bank of 200 with sample size n=5000.*

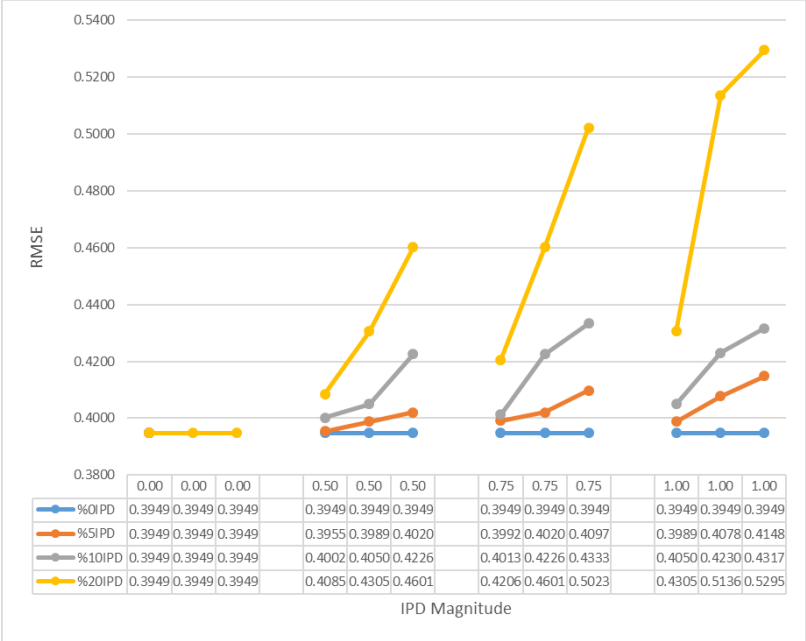


b. *RMSE values for item bank of 500 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951	0.3951
%5IPD	0.3951	0.3951	0.3951	0.3967	0.3994	0.4108	0.3976	0.4108	0.4190	0.3994	0.4097	0.4162
%10IPD	0.3951	0.3951	0.3951	0.3977	0.4134	0.4275	0.4029	0.4275	0.4373	0.4134	0.4291	0.4398
%20IPD	0.3951	0.3951	0.3951	0.4054	0.4447	0.4982	0.4234	0.5049	0.5309	0.4447	0.5309	0.5370

c. *RMSE values for item bank of 1000 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949	0.3949
%5IPD	0.3949	0.3949	0.3949	0.3955	0.3989	0.4020	0.3992	0.4020	0.4097	0.3989	0.4078	0.4148
%10IPD	0.3949	0.3949	0.3949	0.4002	0.4050	0.4226	0.4013	0.4226	0.4333	0.4050	0.4230	0.4317
%20IPD	0.3949	0.3949	0.3949	0.4085	0.4305	0.4601	0.4206	0.4601	0.5023	0.4305	0.5136	0.5295

When RMSE values were examined, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation *RMSE* values obtained at three time points to increase. The increase in the number of measurements in IPD size, IPD percentage, and item banks containing IPD results in more erroneous ability estimations leading to a decrease in measurement precision. Some studies in the literature also show that IPD conditions increase error values (Aksu Dünya, 2017; Babcock & Albano, 2012; Chen, 2013; Risk, 2015; Wells et al., 2002).

When Figure 5. a and c were examined, it is found that the increase in item bank size of 200, 500, and 1000 items resulted in a tendency of RMSE values to decrease. A study by Risk (2015) used item bank sizes of 300, 500, and 1000 and observed that an increase in item bank size

resulted in a decrease in RMSE values. This signifies that an increase in item bank size results in a slight decrease in error values between real and estimated ability values.

Three-factor ANOVA results for independent samples, shown in Table 7, examine whether obtained differences were statistically significant according to above-mentioned RMSE values.

Table 7. *Comparison of RMSE values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=5000.*

Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η2)
200	IPD Size	0.408	3	0.136	7400.06*	0.07
	IPD Percentage	2.760	3	0.920	50059.24*	**0.47**
	Measurement	1.226	2	0.613	22236.46*	0.21
	IPD Size*IPD Percentage	0.267	9	0.030	4842.69*	0.04
	IPD Size*Measurement	0.044	6	0.007	798.05*	0.00
	IPD Percentage*Measurement	0.706	6	0.118	12805.01*	0.12
	IPD Size*IPD Percentage*Measurement	0.085	18	0.005	1541.68*	0.01
	Error	0.262	4752	0.000		
	Total	5.758	4799			
500	IPD Size	0.160	3	0.053	11880.00*	0.03
	IPD Percentage	2.723	3	0.908	202182.75*	**0.53**
	Measurement	1.089	2	0.545	80858.25*	0.21
	IPD Size*IPD Percentage	0.087	9	0.010	6459.75*	0.01
	IPD Size*Measurement	0.151	6	0.025	11211.75*	0.03
	IPD Percentage*Measurement	0.602	6	0.100	44698.50*	0.11
	IPD Size*IPD Percentage*Measurement	0.206	18	0.011	15295.50*	0.04
	Error	0.064	4752	0.000		
	Total	5.082	4799			
1000	IPD Size	0.299	3	0.100	22916.90*	0.08
	IPD Percentage	1.702	3	0.567	130450.06*	**0.48**
	Measurement	0.673	2	0.337	51582.19*	0.19
	IPD Size*IPD Percentage	0.240	9	0.027	18394.84*	0.06
	IPD Size*Measurement	0.063	6	0.011	4828.65*	0.02
	IPD Percentage*Measurement	0.372	6	0.062	28512.00*	0.10
	IPD Size*IPD Percentage*Measurement	0.064	18	0.004	4905.29*	0.02
	Error	0.062	4752	0.000		
	Total	3.475	4799			

**p<.05*

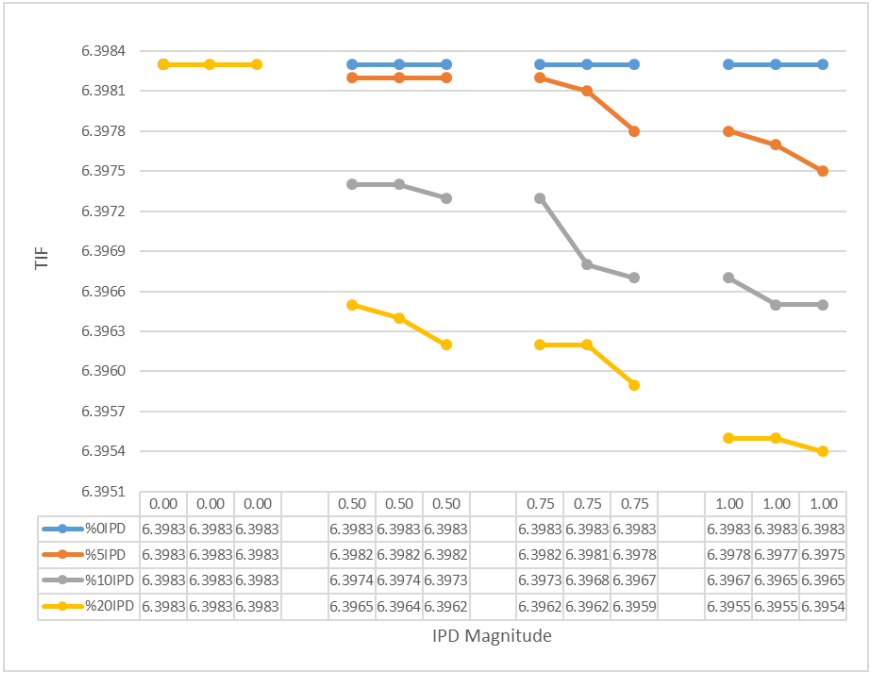
The three-factor ANOVA results for independent samples regarding RMSE values show that the main effect and the effects of the two- and three-way interactions of the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items have statistically significant effects on RMSE. These generally possess low and high effect sizes (Cohen, 1988). The results of post-hoc analysis also revealed differences for every level of every factor. IPD percentage is the factor with the most impact on ability estimation RMSE among the variables within the scope of this study. Some studies in the literature also support this finding (Aksu Dünya; 2017; Babcock & Albano, 2012; Risk, 2015). On the other hand, some studies argue that the impact of IPD on RMSE values was not statistically significant (Chen, 2013; Wells et al., 2002). For instance, Chen (2013) argued that although an increase in the percentage of items containing IPD in the item bank increased RMSE values, this increase was low-level and statistically insignificant.

The findings for TIF values, which constitute the third criterion for comparing independent variable conditions, are shown in Figure 6. a, b and c. When TIF values were examined for a sample size of 5000, the increase in IPD size and IPD percentage for item bank sizes of 200, 500, and 1000 resulted in a tendency of ability estimation *TIF* values obtained at three time points to decrease. The lowest TIF values were obtained for IPD size of 1.00, IPD percentage

of 20, and at the third time point. Therefore, the increase in the number of measurements, IPD size, and IPD percentage results in a decrease in TIF, i.e., the amount of information the test provides for item banks of 200, 500, and 1000. This decreasing tendency does not change with an increase in item bank size. Similarly, TIF values are generally slightly higher in the 5000 sample than 1000 sample, but no increasing or decreasing trend was observed within each sample.

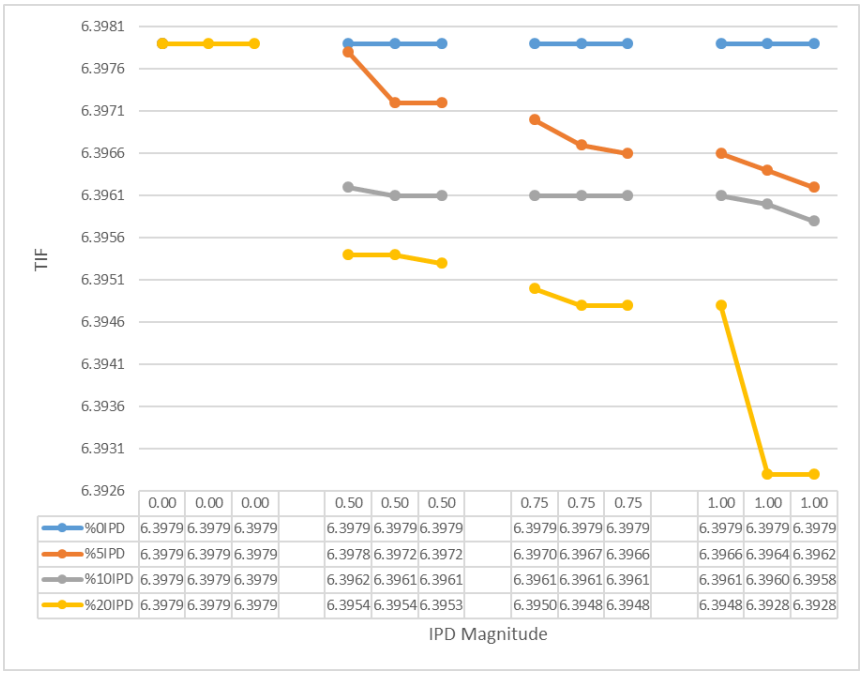
Figure 6. a, b and **c.** *Figures denoting comparison of TIF values at three time points for different item bank sizes with different IPD sizes and different IPD percentages with the sample size is n=5000.*

a. *TIF values for item bank of 200 with sample size n=5000.*



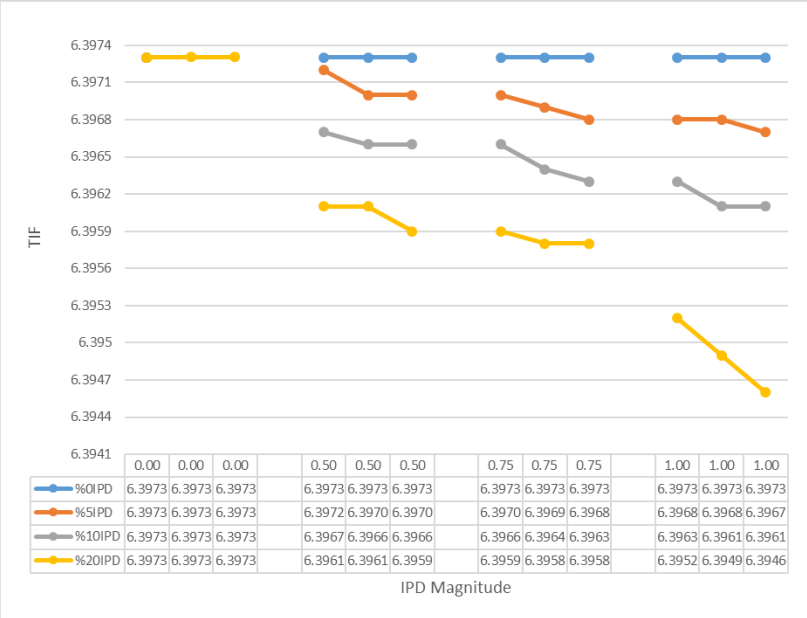
	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983	6.3983
%5IPD	6.3983	6.3983	6.3983	6.3982	6.3982	6.3982	6.3982	6.3981	6.3978	6.3978	6.3977	6.3975
%10IPD	6.3983	6.3983	6.3983	6.3974	6.3974	6.3973	6.3973	6.3968	6.3967	6.3967	6.3965	6.3965
%20IPD	6.3983	6.3983	6.3983	6.3965	6.3964	6.3962	6.3962	6.3962	6.3959	6.3955	6.3955	6.3954

b. *TIF values for item bank of 500 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979	6.3979
%5IPD	6.3979	6.3979	6.3979	6.3978	6.3972	6.3972	6.3970	6.3967	6.3966	6.3966	6.3964	6.3962
%10IPD	6.3979	6.3979	6.3979	6.3962	6.3961	6.3961	6.3961	6.3961	6.3961	6.3961	6.3960	6.3958
%20IPD	6.3979	6.3979	6.3979	6.3954	6.3954	6.3953	6.3950	6.3948	6.3948	6.3948	6.3928	6.3928

c. *TIF values for item bank of 1000 with sample size n=5000.*



	0.00	0.00	0.00	0.50	0.50	0.50	0.75	0.75	0.75	1.00	1.00	1.00
%0IPD	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973	6.3973
%5IPD	6.3973	6.3973	6.3973	6.3972	6.3970	6.3970	6.3970	6.3969	6.3968	6.3968	6.3968	6.3967
%10IPD	6.3973	6.3973	6.3973	6.3967	6.3966	6.3966	6.3966	6.3964	6.3963	6.3963	6.3961	6.3961
%20IPD	6.3973	6.3973	6.3973	6.3961	6.3961	6.3959	6.3959	6.3958	6.3958	6.3952	6.3949	6.3946

Three-factor ANOVA results for independent samples, shown in Table 8, examine whether the differences obtained were statistically significant according to the TIF values discussed above.

Table 8. *Comparison of TIF values according to three time points for different item bank sizes with different IPD sizes and different IPD percentages with sample size n=5000.*

Item Bank Size	Source of Variation	Sum of Squares	*df*	Mean of Squares	F	Effect Size (η2)
200	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.032	3	0.010	38016.00*	**0.45**
	Measurement	0.002	2	0.001	2376.00*	0.03
	IPD Size*IPD Percentage	0.016	9	0.002	19008.00*	0.22
	IPD Size*Measurement	0.007	6	0.001	8316.00	-
	IPD Percentage*Measurement	0.003	6	0.000	3564.00	-
	IPD Size*IPD Percentage*Measurement	0.007	18	0.000	8316.00	-
	Error	0.004	4752	0.000		
	Total	0.071	4799			
500	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.002	3	0.000	2376.00*	0.01
	Measurement	0.000	2	0.000	0.00*	0.00
	IPD Size*IPD Percentage	0.000	9	0.000	0.00*	0.00
	IPD Size*Measurement	0.054	6	0.009	64152.00*	**0.50**
	IPD Percentage*Measurement	0.046	6	0.008	54648.00*	0.43
	IPD Size*IPD Percentage*Measurement	0.000	18	0.000	0.00*	0.00
	Error	0.004	4752	0.000		
	Total	0.106	4799			
1000	IPD Size	0.000	3	0.000	0.00*	0.00
	IPD Percentage	0.001	3	0.000	1584.00*	0.01
	Measurement	0.022	2	0.011	34848.00*	0.20
	IPD Size*IPD Percentage	0.067	9	0.007	106128.00*	**0.63**
	IPD Size*Measurement	0.001	6	0.000	1584.00	-
	IPD Percentage*Measurement	0.003	6	0.001	4752.00	-
	IPD Size*IPD Percentage*Measurement	0.010	18	0.000	15840.00	-
	Error	0.003	4752	0.000		
	Total	0.106	4799			

*$p<.05$

The three-factor ANOVA results for independent samples in terms of TIF values show that both the main effect and the effects of the two- and three-factor interactions of the number of measurements, IPD size, and IPD percentage have statistically significant effects on TIF for an item bank of 500 items and a sample size of 5000. However, these generally possess low effect sizes (Cohen, 1988). IPD size, IPD percentage, measurement and interaction effect of IPD size*IPD percentage for item bank sizes of 200 and 1000 on TIF values were statistically significant (Cohen, 1988). The results of the post-hoc analysis also revealed differences for every level of every factor that revealed significant differences. Therefore, the item bank containing IPD decreases the amount of information the test provides by increasing the errors. Although the TIF values for the 5000 sample size were higher than for the 1000 sample size, the samples themselves show neither an increasing nor a decreasing trend.

4. DISCUSSION and CONCLUSION

This study examines the impact of IPD on measurement precision and TIF in CAT administrations. When the results were examined in terms of measurement precision, the increase in the number of measurements, IPD size, and IPD percentage for item bank sizes of 200, 500 and 1000 items resulted in a decrease in measurement precision because items containing IPD in item bank led to drifts in ability estimations. The increase of IPD in the item bank resulted in bias values growing in the negative direction and RMSE values growing in the positive direction. The cause of positive RMSE values is the square in the RMSE formula. When compared with the baseline data set, the highest values of bias and RMSE were obtained at the third time point, with an IPD size of 1.00 and items containing an IPD percentage of 20. Measurement precision was calculated at its lowest point when conditions for IPD were at the highest point. Three-factor ANOVA for independent samples also revealed statistically significant results regarding these factors for measurement precision and indicated that the factor that affected measurement precision the most was the number of items containing IPD in the item bank. Research findings (Abad et al., 2010; Aksu Dünya, 2017; Babcock & Albano, 2012; Chan et al., 1999; McCoy, 2009; Risk, 2015; Wells et al., 2002) that examine the effects of IPD on measurement precision in CAT administrations are consistent with the finding that argues that the increase in IPD size results in a decrease in precision. While changes in IPD conditions affect measurement precision, an increase in sample size does not result in a changing pattern in either the positive or negative bias direction. The RMSE values were somewhat greater for the 5000-person sample, but no overall growing or declining trend was detected. The study has found that the factor that affected measurement precision the most was IPD percentage. While some studies contend that IPD percentage has the greatest impact on measurement precision (Babcock & Albano, 2012), others contend that IPD size (Risk, 2015), sample size, and IPD percentage (Wells et all, 2002) all influence measurement precision. Using the Rasch model, Risk (2015) examined the effect of various IPD conditions on measurement precision and discovered that the factor affecting measurement precision the most was IPD size rather than the number of items containing IPD in the item bank, but the effect was insignificant. Similarly, Wells et al. (2002) stated in their studies which used the 2PLM model, that sample size and IPD percentage were factors affecting ability estimations the most. It is worth noting that the simulated sample size, item bank size, IPD conditions and the IRT model vary in these studies. While the presence of items containing IPD in the item bank negatively affects measurement precision in CAT administrations, the factor negatively impacts the value depends on the IRT model, sample size, item bank size and IPD conditions.

When the results were examined in terms of TIF values, the increase in the number of items containing IPD in item bank, IPD size and number of measurements in CAT administrations resulted in a slight decrease in the amount of information the test provided. The highest TIF values under all conditions were obtained in the baseline data set not containing IPD, and the

lowest TIF values were obtained at the third time point with the highest rate of IPD conditions. As the number of measurements and IPD conditions increased, the amount of information provided by the test decreased. However, TIF values are generally marginally higher in the 5000-person sample than 1000-person sample, but neither an increasing nor a decreasing trend was observed within each sample. Similarly, there were no observed increasing or decreasing trend in TIF values as the item bank size changed. However, TIF values were affected by the number of measurements and IPD conditions. When the statistical significance of obtained TIF values was examined, statistically significant results were calculated mostly for the main effect and IPD size*IPD percentage factor. While some studies in the literature support the finding of the impact of IPD on TIF values (Chan et al., 1999), other studies obtained statistically insignificant differences (Deng & Melican, 2010; Guo & Wang, 2003). In a study by Guo and Wang (2003), which examined the impacts of parameter drift on CAT using real and simulated data, test characteristic curves were compared. However, since measurements were taken at two time points, very small differences were obtained in terms of TIF values, which were not significant.

In conclusion, this study has found that IPD under-examined conditions negatively affects measurement precision and TIF values. Although the IRT model and CAT administrations bring considerable advantages in ability estimations, the importance of developing tests for the item bank and reviewing items should be particularly emphasized to carry out ability estimations accurately. The chosen way of administrating tests and the models picked for use will only produce accurate results if high-quality items are available in the item bank, and these items can maintain this characteristic.

In light of the study's findings, the following recommendations can be made to researchers: This research was conducted using simulated data. Using test administrations with real data, the impact of IPD on the aforementioned factors could be examined. The examined samples in this study were generated using the normal distribution. However, since non-normally distributed extreme values are frequently encountered in real-world applications, the effects of IPD could also be examined under skewed distribution conditions. This study utilized the Rasch model, and there were no restrictions on item exposure. Consequently, the effects of IPD could also be examined by employing alternative IRT models and imposing various item exposure restrictions. This study examined the conditions under which all individuals may encounter IPD-containing items. However, only a subset of individuals may encounter IPD-containing items due to their prior test-taking experience or a change in the curriculum. Consequently, when IPD-containing items are given to a specific group of individuals in CAT applications, the effects of the condition on ability estimates could be investigated.

Acknowledgments

This paper was produced from the part of the first author's doctoral dissertation prepared under the supervision of the second and third author.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the authors. **Ethics Committee Number**: Ankara University Social Sciences Sub-Ethical Committee, 22/04/2019, 05-181.

Authorship Contribution Statement

Merve Sahin Kursad: Investigation, Methodology, Resources, Visualization, Software, Formal Analysis, and Writing-original draft. **Omay Cokluk Bokeoglu**: Investigation,

Resources, Methodology, Supervision, Writing-original draft. **Rahime Nukhet Cikrikci:** Investigation, Resources, Methodology, Supervision, Writing-original draft.

Orcid

Merve SAHIN KURSAD https://orcid.org/0000-0002-6591-0705
Omay COKLUK BOKEOGLU https://orcid.org/0000-0002-3879-9204
Rahime Nukhet CIKRIKCI https://orcid.org/0000-0003-0876-6644

REFERENCES

Abad, F.J., Olea, J., Aguado, D., Ponsoda, V., & Barrada, J.R. (2010). Deterioro de parámetros de los ítems en tests adaptativos informatizados: estudio con eCAT [Item parameter drift in computerized adaptive testing: Study with eCAT]. *Psicothema, 22,* 340-7.

Aksu Dünya, B. (2017). *Item parameter drift in computer adaptive testing due to lack of content knowledge within sub-populations* [Doctoral dissertation, University of Illinois].

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

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

Barrada, J.R., Olea, J., Ponsoda, V., & Abad, F.J. (2010). A method for the comparison of item selection rules in computerized adaptive testing. *Applied Psychological Measurement, 34*, 438-452. https://doi.org/10.1177/0146621610370152

Bergstrom, B.A., Stahl, J., & Netzky, B.A. (2001, April). *Factors that influence parameter drift* [Conference presentation] American Educational Research Association, Seattle, WA.

Blais, J. & Raiche, G. (2002, April). Features of the sampling distribution of the ability estimate in computerized adaptive testing according to two stopping rules, *International Objective Measurement Workshop*, New Orleans.

Bock, D.B., Muraki, E., & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement, 25*(4), 275-285. https://doi.org/10.1111/j.1745-3984.1988.tb00308.x

Burton, A., Altman, D.G., Royston, P., & Holder, R.L. (2006). The design of simulation studies in medical statistics. *Statistics in Medicine, 25*, 4279-4292. https://doi.org/10.1002/sim.2673

Chan, K.Y., Drasgow, F., & Sawin, L.L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*(4), 610-619. https://doi.org/10.1037/0021-9010.84.4.610

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

Chang, S.W., & Ansley, T.N. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*, 71-103. https://www.jstor.org/stable/1435055

Chen, S.Y., Ankenmann, R.D., & Chang, H.H. (2000). A comparison of item selection rules at the early stages of computerized adaptive testing. *Applied Psychological Measurement, 24*, 241-255. https://doi.org/10.1177/01466210022031705

Chen, Q. (2013). *Remove or keep: linking ıtems showing ıtem parameter drift* [Unpublished Doctoral Dissertation]. Michigan State University.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum.

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

Deng, H., Ansley, T., & Chang, H. (2010). Stratified and maximum ınformation ıtem selection procedures in computer adaptive testing. *Journal of Educational Measurement, 47*(2), 202-226. https://doi.org/10.1111/j.1745-3984.2010.00109.x

Deng, H., & Melican, G. (2010, April). *An investigation of scale drift in computer adaptive test* [Conference presentation] Annual Meeting of National Council on Measurement in Education, San Diego, CA.

Donoghue, J.R., & Isham, S.P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*(1), 33-51. https://doi.org/10.1177/01466216980221002

Eggen, T.J.H.M. (1999). Item selection in adaptive testing with the sequential probability ratio test. *Applied Psychological Measurement*, *23*(3), 249-261. https://doi.org/10.1177/01466219922031365

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

Evans, J.J. (2010). *Comparability of examinee proficiency scores on Computer Adaptive Tests using real and simulated data* [Unpublished Doctoral dissertation]. The State University of New Jersey.

Filho, N.H., Machado, W.L., & Damasio, B.F. (2014). Effects of statistical models and items difficulties on making trait-level inferences: A simulation study. *Psicologia Reflexão e Crítica, 27*(4). https://doi.org/10.1590/1678-7153.201427407

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20*(4), 369-377. https://doi.org/10.1111/j.1745-3984.1983.tb00214.x

Guo, F., & Wang, L. (2003, April). *Online calibration and scale stability of a CAT program* Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Chicago, IL.

Hagge, S., Woo, A., & Dickison, P. (2011, October). *Impact of item drift on candidate ability estimation* [Conference presentation] The Annual Conference of the International Association for Computerized Adaptive Testing, Pacific Grove, CA.

Han, K.T., & Guo, F. (2011). *Potential impact of item parameter drift due to practice and curriculum change on item calibration in computerized adaptive testing* (R-11-02). Graduate Management Admission Council Research Report.

Hatfield, J.P., & Nhouyvanisvong, A. (2005, April). *Parameter drift in a high-stakes computer adaptive licensure examination: An analysis of anchor items* [Conference presentation] The Annual Meeting of the American Educational Research Association, Montreal, Canada.

Huang, C., & Shyu, C. (2003, April). *The impact of item parameter drift on equating* [Conference presentation] National Council on Measurement in Education, Chicago, IL.

Jiang, G., Tay, L., & Drasgow, F. (2009). Conspiracies and test compromise: An evaluation of the resistance of test systems to small-scale cheating. *International Journal of Testing*, *9*(4), 283-309. https://doi.org/10.1080/15305050903351901

Jones, P.E., & Smith, R.W. (2006, April) *Item parameter drift in certification exams and its impact on pass-fail decision making* [Conference presentation] National Council of Measurement in Education, San Francisco, CA.

Kalender, İ. (2011). *Effects of different computerized adaptive testing strategies on recovery of ability* [Unpublished Doctoral Dissertation] Middle East Technical University.

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

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

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

Kim, S.H., & Cohen, A.S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement*, *29*, 51–66. https://www.jstor.org/stable/1434776

Kingsbury, G.G., & Wise, S.L. (2011). Creating a K-12 adaptive test: Examining the stability of ıtem parameter estimates and measurement scales. *Journal of Applied Testing Technology, 12.*

Kingsbury, G.G., & Zara, A.R. (2009). Procedures for selecting items for computerized adaptive tests. *Applied Measurement in Education, 2*(4), 359-375. https://doi.org/10.1207/s15324818ame0204_6

Köse, İ.A. & Başaran, İ. (2021). 2 parametreli lojistik modelde normal dağılım ihlalinin madde parametre kestirimine etkisinin incelenmesi [Investigation of the effect of different ability distributions on item parameter estimation under two-parameter logistics model]. *Journal of Digital Measurement and Evaluation Research*, *1*(1), 01-21. https://doi.org/10.29329/dmer.2021.285.1

Li, X. (2008). An investigation of the item parameter drift in the examination for the certificate of proficiency in English (ECPE). *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6*, 1–28.

Linda, T. (1996, April). *A comparison of the traditional maximum information method and the global information method in CAT item selection* [Conference presentation] National Council on Measurement in Education, New York.

Linden, W.J., & Glas, G.A.W. (2002). *Computerized adaptive testing: Theory and practice.* Kluwer Academic Publishers.

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

McCoy, K.M. (2009). *The impact of item parameter drift on examinee ability measures in a computer adaptive environment* [Unpublished Doctoral Dissertation]. University of Illinois.

McDonald, P.L. (2002). *Computer adaptive test for measuring personality factors using item response theory* [Unpublished Doctoral Dissertation]. The University Western of Ontario.

Meng, H., Steinkamp, S., & Matthews-Lopez, J. (2010). *An investigation of item parameter drift in computer adaptive testing* [Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Denver, CO.

Meyers, J., Miller, G.E., & Way, W.D. (2009, April). *Item position and item difficulty change in an IRT based common item equating design* [Conference presentation] The American Educational Research Association, San Francisco, CA.

Nydick, S.W. (2015). An R package for simulating IRT-based computerized adaptive tests.

Patton, J.M., Cheng, Y., Yuan, K.H., & Diao (2013). The influence of item calibration error on variable-length computerized adaptive testing. *Applied Psychological Measurement, 37*(1), 24–40. https://doi.org/10.1177/0146621612461727

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

Reckase, M.D. (2011). Computerized adaptive assessment (CAA): The way forward. In *The road ahead for state assessments, policy analysis for California education and Rennie Center for Education Research & Policy* (pp.1-11). Rennie Center for Education Research & Policy.

Risk, N.M. (2015). *The impact of item parameter drift in computer adaptive testing* (CAT) [Unpublished Doctoral Dissertation]. University of Illinois.

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

Rupp, A.A., & Zumbo, B.D. (2003). *Bias coefficients for lack of invariance in unidimensional IRT models.* Vancouver: University of British Columbia.

Rupp, A.A., & Zumbo, B.D. (2004). A note on how to quantify and report whether item parameter invariance holds: When Pearson correlations are not enough. *Educational and Psychological Measurement*, *64*, 588-599. https://doi.org/10.1177/0013164403261051

Schulz, W., & Fraillon, J. (2009, September). The analysis of measurement equivalence in ınternational studies using the rasch model [Conference presentation] The European Conference on Educational Research (ECER), Vienna.

Scullard, M.G. (2007). *Application of item response theory based computerized adaptive testing to the strong interest inventory* [Unpublished Doctoral Dissertation]. University of Minnesota.

Segall, D.O. (2004). Computerized adaptive testing. In K. Kempf-Lenard (Ed.), *The Encyclopedia of social measurement*. Academic Press.

Song, T., & Arce-Ferrer, A. (2009, April). *Comparing IPD detection approaches in common-item nonequivalent group equating design* [Conference presentation] The Annual Meeting of the National Council on Measurement, San Diego, CA.

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

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

Svetina, D., Crawford, A.V., Levy, R., Green, S.B., Scott, L., Thompson, M., Gorin, J.S., Fay, D., & Kunze, K.L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-PL. *Psychological Test and Assessment Modeling, 55*(4), 335-360.

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

Veldkamp, B.P., & Linden van der, W. (2006) Designing item pool for computerized adaptive testing. In *Designing Item Pools* (pp.149-166). University of Twente.

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

Wainer, H., Dorans, N.J., Eignor, D., Flaugher, R., Green, B.F., Mislevy, R.J., Steinberg, L., & Thissen, D. (2010). *Computerized adaptive testing: A primer*. Lawrence Erlbaum Associates Publishers.

Wang, T. (1997, March). *Essential unbiased EAP estimates in computerized adaptive testing* [Conference presentation] The American Educational Association, Chicago, IL.

Wang, H-P., Kuo, B-C., Tsai, Y-H., & Liao, C-H. (2012). A Cerf-Based computerized testing system for Chinese proficiency. *TOJET: The Turkish Journal of Educational Technology, 11*(4), 1–12.

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

Wells, C.S., Subkoviak, M.J., & Serlin, R.C. (2002). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*(1), 77-87. https://doi.org/10.1177/0146621602261005

Wise, S.L., & Kingsbury, G.G. (2000). Practical issues in developing and maintaining a computerized adaptive testing program. *Psicológica, 21*(2000), 135-155.

Witt, E.A., Stahl, J.A., Bergstrom, B.A., & Muckle, T. (2003, April). *Impact of item drift with nonnormal distributions* [Conference presentation] The Annual Meeting of the American Educational Research Association, Chicago, IL.

Wollack, J.A., Sung, H.J., & Kang, T. (2005) *Longitudinal effects of item parameter drift* [Conference presentation] The Annual Meeting of the National Council on Measurement in Education, Montreal, CA.

Yao, L. (2013). Comparing the performance of five multidimensional CAT selection procedures with different stopping rules. *Applied Psychological Measurement, 37*(1), 3-23. https://doi.org/10.1177/0146621612455687

Yi, Q., Wang, T., & Ban, J.C. (2001). Effects of scale transformation and test termination rule on the precision of ability estimation in computerized adaptive testing. *Journal of Educational Measurement, 38,* 267-292. https://doi.org/10.1111/j.17453984.2001.tb01127.x