The Estimation Stability of Logistics Model Parameters Under Different Scenarios

Andhita Dessy Wulansari[12]

[1]Institut Agama Islam Negeri Ponorogo, Ponorogo 63471, Indonesia,
(email: andhita@iainponorogo.ac.id)

[2]Department of Educational Research and Evaluation, Graduate School of Yogyakarta State University, Yogyakarta 55281, Indonesia,
(email: andhita.dessy2015@student.uny.ac.id)

Abstract
This research identifies how the contribution of IRT model, sample size N and number of item/test length n to the stability estimation of item parameter and examinee parameter Θ. The data used in this study is generated by WINGEN, and the parameter estimation uses BILOG-MG. The trend pattern is observed based on the coefficiency correlation between the actual parameter and the estimated parameter. The findings of this research are: (1) the greater the N, the stabler the item parameter; (2) the more n, the stabler the estimation of examinee parameter Θ; (3) one-parameter logistic model is the stablest model in estimating item b parameter, two-parameter logistic model is the stablest model in estimating parameter item a and three-parameter logistic model is the stablest model in estimating item parameter c, whereas two-parameter logistic model is the stablest model in estimating examinee parameter (4) the stability of the item parameter is more affected by the sample size N than the number of item/test length n, while the stability of examinee parameter Θ is more influenced by the number of item/test length n than sample size N; (5) In one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model, the stability of item parameter is more influenced by sample size N than the number of item/test length n while the stability of examinee parameter Θ is more influenced by the number of item/test length n than sample size N, where one-parameter logistic model is the best model in estimating item parameter and two-parameter logistic model is the best model in estimating examinee parameter, and three-parameter logistic model is the most unstable model in estimating both item and examinee parameter.

Keywords: IRT model, sample size, test length, item parameter, examinee parameter

INTRODUCTION

In the field of measurement, the effect of the Item Response Theory (IRT) model, the sample size N and the number of item/test length n on the stability estimation of item parameter and examinee parameter Θ are often questioned as they are related to the logistics model accuracy of the IRT. This was first discussed by Hambleton & Cook (1983) in a study about the reliability of IRT model and the effect of test length n and sample size N on the accuracy of the examinee parameter estimation. According to the study, test length n and sample size N are important factors that can influence the accuracy of estimation curve of measurement error. In addition, Suwarto (2005), who examined the effect of sample size N and logistic model on item parameter estimation, concluded that for the purposes of testing and collecting data using instruments (tests) where the calculation use IRT, it needs a large sample size N. One-parameter logistic model is the most suitable for item b parameter estimation, two-parameter logistic model is the most suitable for item a parameter estimation, and three-parameter logistic model is the most suitable for item c parameter estimation. Retnawati (2006) also studied the stability of parameter estimation on logistic regression. The results of this study indicate that based on the results of the analysis of significance using variance analysis, distribution ability, length of test and interaction of test length with sample size affect the stability of parameter estimation item b only. Sahin & Anil (2017) examined the effect of test length n and sample size N on the item parameter estimation. The results of this study indicate that the combination of test length n and sample size N is very important to consider as it affects the accuracy of IRT model.

The purpose of research is wider than those mentioned above, that is to identify: (1) the effect of sample size N on the stability of item parameter estimation; (2) the influence of number of item/test length n on the stability of examinee parameter ɵ; (3) the effect of the model on the stability of item parameter estimation and examinee parameter ɵ; (4) the effect of sample size n and number of item/test length n on the stability of item parameter estimation and examinee parameter ɵ; (5) the influence of model, sample size n and number of item/test length n on the estimation stability of parameter item and examinee parameter ɵ.

LITERATURE REVIEW

Measurement

A measurement must be done on purpose. In the world of education, a set of instruments is arranged to determine the characteristics of the object of the study that can be school or university students, teachers, or lecturers. To obtain accurate information, it needs a good measurement tool. According to Cronbach (1990), the obligatory measurement instruments in education must have validity and reliability. Validity can be interpreted as to what extent the measurement instrument can perform its functions appropriately. According to Crocker & Algina (1986), validity includes content, construction and

criteria. Reliability can be interpreted as to what extent the measurement results of the measurement instrument can be trusted, i.e. the results obtained from several tests for the same group of objects are relatively similar.

In Classical Test Theory (CTT), measurement has unique characteristics. In concluding the CTT, the group of items or questionnaire are dependent on the group of test participants who respond. If the questionnaire is responded by a different group of test participants, then the characteristics of the item such as difficulty level (b) will change. If the questionnaire is responded by a group of highly skilled participants, the difficulty level will be low and vice versa. Similar to the characteristics of items, the ability of the test group (Θ) on CTT will change if different questionnaires are given. The group of test takers working on the easy questionnaire will appear to be highly capable and the group of test takers working on the difficult questionnaire will appear to be low-skilled. Based on this explanation, it can be seen that on CTT, information about characteristics of item cannot be used to determine the ability of test participants accurately. Whereas in the measurement, the relation between the answers of the test participants and the items can produce results that should be accurate, so the conclusion can be used as appropriate information.

To overcome dependencies between groups of test participants and the questionnaires on CTT, the Item Response Theory (IRT) is introduced. The basic principle of IRT is that the characteristics of the questionnaire will not change even if answered by a group of test participants with different abilities, and the ability of the test takers will not change even if given questionnaires with different characteristics.

Item Response Theory (IRT)

Item Respons Theory (IRT) is presented to overcome the weaknesses in Classical Test Theory (CTT). According to Hambleton & Rogers (1991), the probability of answering the items correctly, the characteristics of the item parameter and the characteristics of the test takers parameter on IRT are connected in an equation model. Using this model, if the difficulty level of the item is identified then the ability of the test participants can be determined and if the ability of the test participants is identified then the difficulty level of the item can be determined. In this equation model, there is invariance (fixed/unchanged) of the characteristics of items and test participants. This invariance distinguishes between IRT and CTT, so that the ability of test participants (Θ) and the difficulty level of the items (b) can be independent. Using the independence of the characteristics of itemss and test participants, the test participants can get the items according to their ability. In IRT model, the equation model can be differentiated based on the number of parameter of the items characteristics, namely one-parameter logistics model, two-parameter logistic model and three-parameter logistic model.

In IRT equation model, a test taker usually has different ability parameter with the other test takers. The test takers' ability parameter is generally symbolized as Θ (theta). According to Linn (1989), the value of Θ can extend from -∞ up to ∞ . However, according to Hambleton, Swaminathan & Rogers (1991), the value of Θ can be determined within a standard interval between -4 and 4.

In one-parameter logistic model, according to Hambleton, Swaminathan & Rogers (1991), the test takers probability in answering the items correctly can be written in the equation model below.

$$p_i(X_{ij}=1|\theta_i,b_j)=\frac{\exp(\theta_i-b_j)}{1+\exp(\theta_i-b_j)} \quad (1)$$

The degree of difficulty in item (b) in one-parameter logistic model can be defined as a parameter showing the scale of the test takers' ability whose probability in answering the item correctly is 0.5 (Hambleton, Swaminathan & Rogers, 1991). The easier the item or the smaller the value of b, the greater the probability of the test participants in answering the item correctly p (Θ) and vice versa. The item is categorized as easy if the test takers' ability is greater than the difficulty level of the item and vice versa. According to Aiken (1994), the purpose of the difficulty level of the item is usually associated with the purpose of the test. Items with high difficulty level are for entry selection test, items with medium difficulty level are for semester exam and items with low difficulty level are for diagnostic test.

In two-parameter logistic model, according to Hambleton, Swaminathan & Rogers (1991), the probability of the test takers in answering the items correctly can be formulated in the following equation.

$$p_i(X_{ij}=1|\theta_i,b_j)=\frac{\exp(a_j(\theta_i-b_j))}{1+\exp(a_j(\theta_i-b_j))} \quad (2)$$

Besides the level of difficulty level parameter (b), two-parameter logistic model also has difference power parameter (a). The existence of difference power parameter (a) in the equation model can be used to identify whether the items analyzed can distinguish the characteristics of the test takers or not. According to Surapranata (2004), the level of difficulty affects the difference power. If every test taker answers correctly or wrongly, the item then cannot distinguish the test takers' ability properly. Items can be categorized as having a good difference power if they can be answered correctly by most the test takers who are clever and cannot be answered correctly by most test takers whose ability is low.

In three-parameter logistic model, according to Hambleton, Swaminathan & Rogers (1991), the probability of the test takers in answering the items correctly can be written in the equation model below.

$$p_i(X_{ij}=1|\theta_i,b_j)=c_i+(1-c_i)\frac{\exp(a_j(\theta_i-b_j))}{1+\exp(a_j(\theta_i-b_j))} \quad (3)$$

Besides the parameter of the items' difficulty level (b), and difference power (a), there is also also pseudo guessing parameter (c) in three-parameter logistic model Pseudo guessing can be defined as the probability of test takers in guessing the answer correctly.

In order for the above three models to produce unbiased parameter estimation, several assumptions have to meet. First, the assumption that the items are only used to measure the ability. This assumption is called unidimensional. For instance, a test device is developed with the aim of measuring the mathematical abilities of the test takers, then the item item should be really just to measure mathematics skills rather than the combination of mathematics and language skills (if math problems are presented in English). But in reality, it is very difficult to obtain items that meet unidimensional assumptions. If the answer of the test takers is the

combination of several abilities, the contribution of each ability to the test takers' answers cannot be identified. This unidimensional assumption is to maintain invariance on IRT. If this assumption is not met, when the group of items is replaced or the group of test takers is replaced, the invariance cannot be maintained. The second is the assumption of local independence, which according to Allen & Yen (2001) can be divided into 2 i.e. local independence of test takers and local independence of the item. Local independence of the test takers is whether the test taker's answer is correct or wrong, it does not affect the other test takers' answer in aswering that same item. Local independence of the test takers can be interpreted as whether the test taker's answer is correct or wrong, it does not affect the test taker's answer in aswering different item.

MATERIAL & METHODOLOGY

To achieve the objectives of the study, it is necessary to simulate a computer program and each case is replicated 5 times. The data is generated by WINGEN, using parameter estimation using BILOG-MG. The computer simulations are run using each of N (32, 288 and 1152). The amount of N is determined based on the assumption that the maximum number of students in each class of Junior High School is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, then the number of students is 1152. While each of the n is 20, 40 and 80. The amount of n=40 is determined based on the assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Along with those assumptions, the size of the sample size N and the test length n are also determined by considering some previous studies. For one-parameter logistic model, according to Goldman & Raju (1986), the minimum sample size is 250, according to Guyer & Thompson (2012), the minimum sample size is 300, and according to Thissen & Wainer (1982); Stone (1992) the minimum sample size is 500. In addition, Harwell & Janosky (1991) combined sample size of 250 with test length of 15 items, Lim & Drasgow (1990) combined sample size of of 750 with test length 20 items, Weiss & Minden (2012) combined a sample size of 200 with a test length of 25 items, Hulin, Lissak, & Drasgow combined a sample size of 500 with a test length of 30 items, Yoes (1995) combined a sample size of 300 with a test length of 75 items. For two-parameter logistic model, Sahin & Anil (2017) combined sample sizes of 150, 250, 350, 500, and 750 with test lengths of 10, 20, and 30 items. For three-parameter logistic model, Patsula & Gessaroli (1995); Swaminathan & Gifford (1983); Yen(1987) combined a sample size of 1000 with a test length of 20 items, Weiss & Minden combined sample size 200 with a test length of 25 items, Akour & Al-Omari (2013) combined a sample size of 500 with a test length of 30 items, Chuah, Drasgow, & Luecht (2006) combined a sample size of 300 with a test length of 50 items, Patsula & Gessaroli (1995); Tang et al. (1993); Yen (1987) combined a sample size of 1000 with a test length of 40 items, Lord (1967) combined a sample size of 1000 with a test length of 50 items, Hulin (1982) combined a sample size of 1000 with a test length of 60 items, Yoes (1995) combined a sample size of 1000 with test length 75, Ree & Jensen (1983) combined a sample size of 500 with a test length of 80 items.

The results of this simulation are then compared between true parameter (results of WINGEN program) and item parameter estimation (results of BILOG-MG program) by using correlation method in Excel program. The criterion used to determine the stability of the parameter is the average correlation between true parameter and estimated parameter. That which is closest to one is the most stable.

RESULTS AND DISCUSSION

In this research, the effect of Item Response Theory (IRT) model, sample size N and number of item/test length n on the stability estimation of item parameter and examinee parameter Θ will be analyzed.

A. The Effect of Sample Size N on Estimation Stability of Item Parameter

To determine the effect of sample size N on the stability of item parameter estimation, it can be seen from the size ratio of the RMSE between the number of sample size N or the ratio of correlation between the number of sample size N. The following data is the correlation between true parameter and estimated parameter item (can only be seen in column b because the model used is one-parameter logistic model, considering that all models have item b parameter) with 5 times replication, for each N (32, 288 and 1152). The number of N is determined based on the assumptions that the maximum number of students in each class in junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, then the total number of students is 1152.

Table I. Correlation between True Parameter and Estimated Parameter Item (Sample Size Factor N)

	Correlation coefficients (N=32)			*Correlation coefficients (N=288)*			*Correlation coefficients (N=1152)*		
	A	b	c	A	B	C	a	b	c
Rep1	0	0,933	0	0	0,992	0	0	0,999	0
Rep2	0	0,952	0	0	0,994	0	0	0,998	0
Rep3	0	0,925	0	0	0,995	0	0	0,998	0
Rep4	0	0,937	0	0	0,995	0	0	0,998	0
Rep5	0	0,918	0	0	0,99	0	0	0,999	0

From each of N (32, 288 and 1152), the average value of correlation between true parameter and estimated parameter item with 5 times replication is as follows:

Table II. Mean Correlation between True Parameter and Estimated Parameter Item (Sample Size Factor N)

	AVERAGE CORRELATION
N32	0,933
N288	0,9932
N1152	0,9984

Based on the tables above, the correlation graph vs sample size (N) can be illustrated in the following figure.

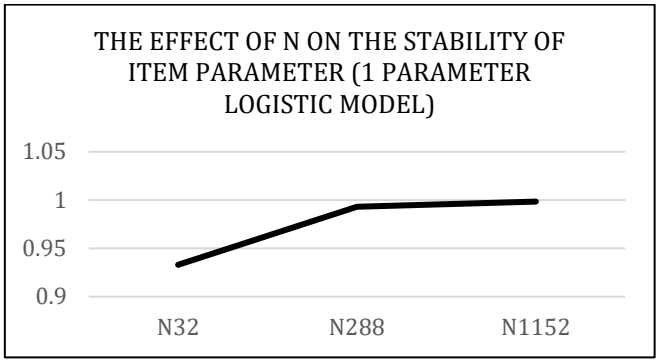

Fig 1. The Effect of N on the Stability of Item Parameter (One-parameter Logistic Model)

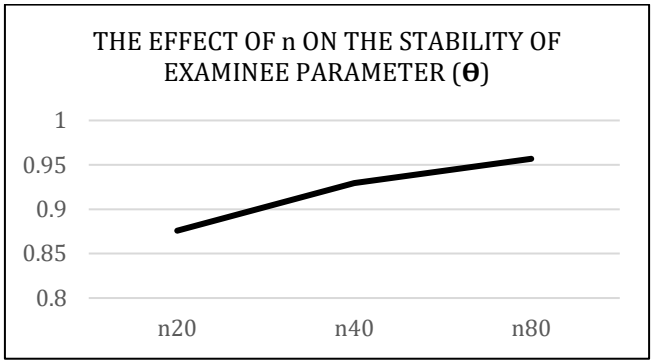

Fig 2. The Effect of n on the Stability of Examinee Parameter

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that N1152 has the highest correlation and N32 has the lowest correlation. Although the difference between the number of N is quite small, it can be concluded that the greater the N, the higher the the correlation value between true parameter with estimate parameter items (closer to 1), so that it can be interpreted that the greater the N, the stabler the item parameter estimation.

B. The Effect of Number of Item/Test Length n on the Stability of Examinee Parameter Θ

To determine the effect of number of item/test length n on the stability of examinee parameter estimation (Θ), it can be seen from comparison of correlation between number of item/test length n. The following data is the correlation between true parameter and estimated examinee parameter (Θ) with replication of 5 times, for each n (20, 40 and 80). The number of n is determined based on assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Table III. The Correlation between True Parameter and Estimated Item Parameter (Factor Number of Item/Test Length n)

	Correlation coefficients (n=20)	*Correlation coefficients (n=40)*	*Correlation coefficients (n=80)*
	true vs estimate	*true vs estimate*	*true vs estimate*
Rep1	0.870388635	0.926581973	0.957881207
Rep2	0.880314137	0.928198813	0.956740163
Rep3	0.882142040	0.933488851	0.954682311
Rep4	0.876159811	0.929538179	0.958529085
Rep5	0.870130316	0.929886287	0.956322850

From each n (20, 40 and 80), the average value of correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication is as follows:

Table IV. The Average Correlation between True Parameter and Estimated Item Parameter (Factor Number of Item/Test Length n)

	AVERAGE CORRELATION
n20	0.875827
n40	0.929539
n80	0.956831

Based on the tables above, the correlation vs. test length (n) can be illustrated in the following graph.

What is addressed as the stability criterion is the greater the coelation, the stabler the estimated parameter. Based on the graph above, it can be interpreted that n 80 has the highest correlation and n 20 has the lowest correlation. Although the correlation between n is quite small, it can be concluded that the more n, the higher the correlation between true parameter and estimated examinee parameter (Θ) or it can be interpreted that the more n, the stabler the examinee parameter (Θ) .

C. The Effect of Model on Stability Estimation of Item Parameter and Examinee Parameter Θ

The effect of model on the stability of item parameter estimation and examinee parameter estimation (Θ) can be seen from comparison of correlation between models.

Based on WGZ File output, the correlation between true parameter and estimated item a, b, c parameter with 5 timesreplication for each model (one-parameter logistic model, two-parameter logistic model and three-parameter logistic model) can be obtained. From each model, the average value of correlation between true parameter and estimated item parameter is as follows.

Table V. The Correlation between True Parameter and Estimated Item Parameter (Model Factor)

	AVERAGE CORRELATION (Parameter item b)			*AVERAGE CORRELATION (Parameter item a)*			*AVERAGE CORRELATION (Parameter item c)*		
	NPARM=1	NPARM=2	NPARM=3	NPARM=1	NPARM=2	NPARM=3	NPARM=1	NPARM=2	NPARM=3
1PL	**0.9972**	0.9948	0.9912	0	0	0	0	0	0
2PL	0.9698	0.9944	0.967	0.00667	**0.932**	0.882	0	0	0
3PL	0.9534	0.9658	0.981	0.001333	0.6332	0.8174	0	0	**0.5236**

What is menat by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the above table, it can be interpreted that to estimate item b parameter, it is better to use one-parameter logistic model, as two-parameter logistic model and three-parameter logistic model are not as stable as one-parameter logistic model in estimating item b parameter. This is indicated by the high correlation on one-parameter logistic cell vs NPARM=1, that is 0.9972. To estimate the parameter of item a, it is better to use two-parameter logistic model because by using the one-parameter logistic model, all the same correlation is 0, whereas three-parameter logistic model is not as stable as two-parameter logistic model in estimating item a parameter. This is indicated by the high correlation on two-parameter logistic cell vs NPARM=2, that is 0.932. To estimate item c parameter, it is better to use three-parameter logistics model because the one-parameter logistic model and two-parameter logistic model assume no guessing factor. The guessing factor is equal to 0, so the low-ability person is assumed that c=0.

With the assistance of SPSS software, the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication for each model (one-parameter

logistic model, two-parameter logistic model and three-parameter logistic model) can be obtained. From each model, the average value of correlation between true parameter and estimated examinee parameter is as Table VI.

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the table above, it can be interpreted that two-parameter logistic model have the highest average correlation (close to 1), and the average correlation between true parameter and estimated examinee parameter on two-parameter logistic model shows that two-parameter logistics model is the stablest. In this case, based on the correlation, the two-parameter logistics model is the best model to estimate the examinee parameter.

Table VI. The Average Correlation between True Parameter and Estimated Item Parameter (Model Factor)

	AVERAGE CORRELATION		
	NPARM=1	NPARM=2	NPARM=3
One-parameter logistic model	0.8983	0.835676	0.895972
Two-parameter logistic model	0.902482	0.911071	0.90922
Three-parameter logistic model	0.87084	0.886444	0.887984

D. The Effect of Sample Size N and Number Of Item/Test Length n on the Estimation Stability of Item Parameter and Examinee Parameter Ɵ.

To determine the effect of sample size N and number of item/test length n on the estimation stability of item parameter and examinee parameter (Ɵ), it can be seen from the comparison of correlation between sample size N and comparison of correlation between number of item/test length n.

The following data is the correlation between true parameter and estimated item parameter (can only be seen in column b because the model used is one-parameter logistic model considering that other models have item b parameter) with 5 times replication, for each N (32, 288 and 1152) and n (20, 40 and 80). The number of N is determined based on the assumptions of the maximum number of students in each class in one junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions of the number of items in the National Exam of Junior High School which amounted to 40 items, and the researcher intends to identify how the difference of correlation between true parameter and estimated examinee parameter (Ɵ) if the item number is shortened to ½ time that is 20 and if it is lengthened twice, which is 80.

Table VII. The Correlation between True Parameter and Estimated Item Parameter (Sample Size Factor N 32 and Number of Item/Test Length n)

	Correlation coefficients (N=32; n=20)			*Correlation coefficients (N=32; n=40)*			*Correlation coefficients (N=32; n=80)*		
	a	b	c	a	b	c	A	b	c
Rep1	0	0.973	0	0	0.933	0	0	0.923	0
Rep2	0	0.946	0	0	0.952	0	0	0.921	0
Rep3	0	0.958	0	0	0.925	0	0	0.931	0
Rep4	0	0.93	0	0	0.937	0	0	0.926	0
Rep5	0	0.95	0	0	0.918	0	0	0.941	0

Table VIII. The Correlation between True Parameter and Estimated Item Parameter Item (Sample Size Factor N 288 and Number of Item/Test Length n)

	Correlation coefficients (N=288; n=20)			*Correlation coefficients (N=288; n=40)*			*Correlation coefficients (N=288; n=80)*		
	a	b	c	A	b	c	A	b	c
Rep1	0	0.904	0	0	0.992	0	0	0.992	0
Rep2	0	0.943	0	0	0.994	0	0	0.996	0
Rep3	0	0.884	0	0	0.995	0	0	0.992	0
Rep4	0	0.953	0	0	0.995	0	0	0.99	0
Rep5	0	0.964	0	0	0.99	0	0	0.991	0

Table IX. The Correlation between True Parameter and Estimated Item Parameter (Sample Size Factor N 1152 and Number of Item/Test Length n)

	Correlation coefficients (N=1152; n=20)			*Correlation coefficients (N=1152; n=40)*			*Correlation coefficients (N=1152; n=80)*		
	a	b	c	A	b	c	A	b	c
Rep1	0	0.998	0	0	0.999	0	0	0.998	0
Rep2	0	0.998	0	0	0.998	0	0	0.998	0
Rep3	0	0.998	0	0	0.998	0	0	0.998	0
Rep4	0	0.999	0	0	0.998	0	0	0.998	0
Rep5	0	0.999	0	0	0.999	0	0	0.998	0

From each N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated item parameter obtained is presented in the following table.

Table X. The Mean Correlation between True Parameter and Estimated Item Parameter (Sample Size N Factor and Number of Item/Test Length n)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.9514	0.933	0.9284
N288	0.9296	0.9932	0.9922
N1152	0.9984	0.9984	0.998

Based on the table above, the correlation graph vs sample size (N) and test length (n) can be illustrated in the following figure.

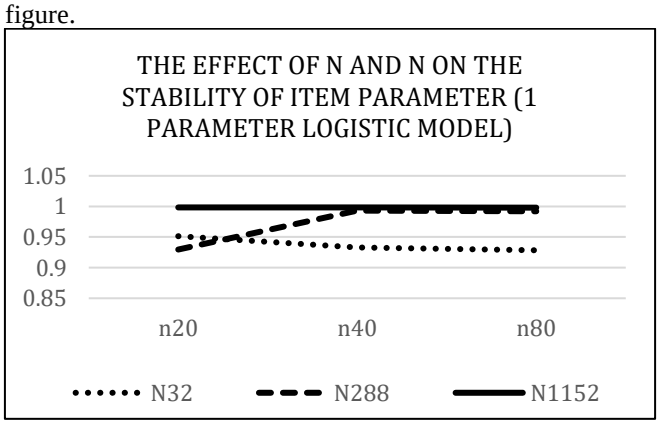

Fig 3. The Effect of N and n on the Stability of Item Parameter (One-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that N32 has the lowest correlation and N1152 has the highest correlation (can be seen from its line position which is approaching 1). As for the value of n in N1152, all three have correlation values that have very little different or tend to be stable at the same number. This shows that the stability of item parameter is

more influenced by the number of sample size N than the number of item/test length n. In general, it can be concluded that the number of N gives a linear effect on the stability of item parameter estimation. The greater the N, the higher the correlation between true parameter and estimated item parameter or it can be said that, the greater the N, the stabler the item parameter estimation.

The following data is the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication, for each N (32, 288 and 1152) and n (20, 40 and 80). The number of N is determined based on the assumptions that the maximum number of students in each class in junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items of National Exam in Junior High School is amounted to 40 items, and the researcher wants to identify the difference of the correlation between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, which is 80.

Table XI. The Correlation between True Parameter and Estimated Examinee Parameter (Sample Size Factor N 32 and Number of Item/Test Length n)

	Correlation coefficients (N32-n20)	*Correlation coefficients (N32-n40)*	*Correlation coefficients (N32-n80)*
	true vs estimate	*true vs estimate*	*true vs estimate*
Rep1	0.897806364	0.966630479	0.965284792
Rep2	0.857208734	0.917374747	0.955700612
Rep3	0.844791195	0.940501252	0.969398315
Rep4	0.940122342	0.941407736	0.965382344
Rep5	0.845681664	0.906516628	0.956562226

Table XII. The Correlation between True Parameter and Estimated Examinee Parameter (Sample Size Factor N 288 and Number of Item/Test Length n)

	Correlation coefficients (N288-n20)	*Correlation coefficients (N288-n40)*	*Correlation coefficients (N288-n80)*
	true vs estimate	*true vs estimate*	*true vs estimate*
Rep1	0.859443166	0.924262578	0.956287517
Rep2	0.879468274	0.937472843	0.956822987
Rep3	0.881716129	0.921142534	0.955145014
Rep4	0.874214193	0.932264624	0.9561023
Rep5	0.887672826	0.928628824	0.955559912

Table XIII. The Correlation between True Parameter and Estimated Examinee Parameter (Sample Size Factor N 1152 and Number of Item/Test Length n)

	Correlation coefficients (N1152-n20)	*Correlation coefficients (N1152-n40)*	*Correlation coefficients (N1152-n80)*
	true vs estimate	*true vs estimate*	*true vs estimate*
Rep1	0.869186935	0.926911074	0.957816075
Rep2	0.872665592	0.934928556	0.95521846
Rep3	0.87901515	0.930086731	0.958038157
Rep4	0.878707352	0.930740729	0.956178441
Rep5	0.873089668	0.927885253	0.956495613

From each N (32, 288 and 1152) and n (20, 40 and 80) the average value of correlation between true parameter and estimated examinee parameter (Θ) obtained is as follows.

Table XIV. The Mean Correlation between True Parameter and Estimated Examinee Parameter (Sample Size N Factor and Number of Item/Test Length n)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.877122	0.934486	0.962466
N288	0.876503	0.928754	0.955984
N1152	0.874533	0.93011	0.956749

Based on the table above, the correlation graph vs sample size (N) and test length (n) can be illustrated in the following figure.

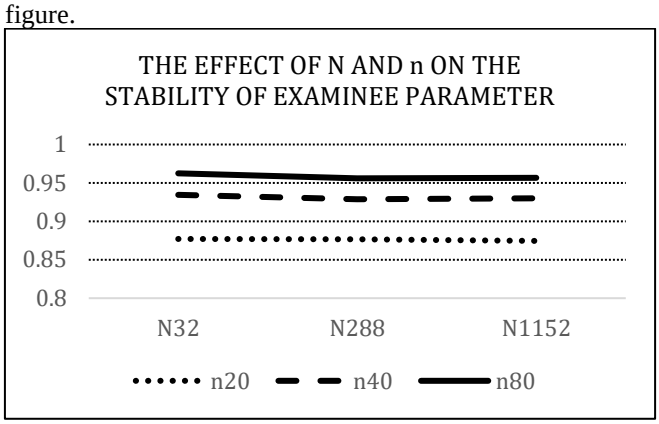

Fig 4. The Effect of N and n on the Stability of Examinee Parameter

What is meant by the stability criterion here is, the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that n 20 has the lowest correlation and n 80 has the highest correlation (can be seen from the position of the line that is approaching 1). As for the N value at n 80, all three have correlation values that have little difference or tend to be stable at the same number). This indicates that the stability of the examinee parameter (Θ) is more influenced by the number of item/test length n than the sample size N. Thus, it can be concluded that the number of n gives a linear effect on the stability of examinee parameter estimation (Θ). The greater the n, the higher the correlation value between true parameter and estimated examinee parameter (Θ) or it can be interpreted that, the bigger the n, the stabler the examinee parameter (Θ).

E. The Effect of Model, Sample Size N and Number of Item/Test Length n to the Stability Estimation of Item Parameter and Examinee Parameter Θ

To determine the effect of model, sample size N and number of item/test length n on the estimation stability of item parameter and examinee parameter (Θ), it can be seen from comparison of correlation between sample size N and comparison of correlation between number of item/test length n on each model of one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model

1. One-Parameter Logistic Model

Using the same procedures as those of Problem No. 4, the correlation between true parameter and estimated item parameter (can only be seen in column b because the model used is one-parameter logistic model) with 5 times replication, for each N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in each class in junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The n is determined based on assumptions that the total number of items in National Exam in Junior High School is 40 items, and the researcher wants to identify the difference between the correlation between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated item parameter obtained is as follows.

Table XV. The Average Correlation between True Parameter and Estimated Item b Parameter (One-Parameter Logistics Model)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.9514	0.933	0.9284
N288	0.9296	0.9932	0.9922
N1152	0.9984	0.9984	0.998

Based on the table above, the correlation graph vs sample size (N) and test length (n) can be illustrated in the following figure.

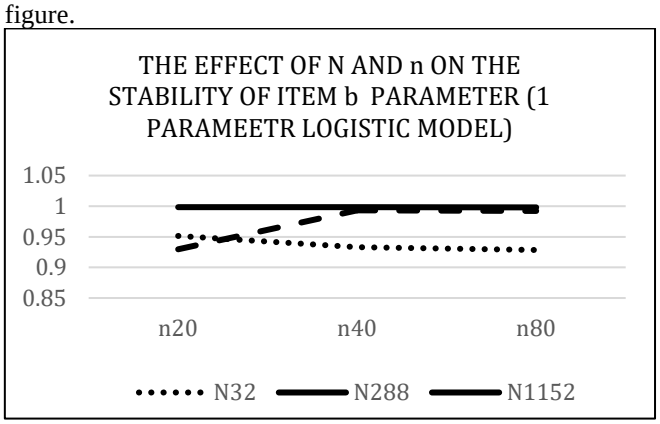

Fig 5. The Effect of N and n on the Stability of Item b Parameter (One-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph of parameter b above, it can be interpreted that N32 has the lowest correlation line and N1152 has the highest correlation line (can be seen from its line position which is approaching 1). As for the value of n in N1152, all three have correlation values have little different or tend to be stable at the same number. This shows that in one-parameter logistic model, the stability of item parameter is more influenced by sample size N than the number of item/test length n. In general, it can be concluded that the number of N gives a linear effect on the stability of item parameter estimation. The greater the N, the higher the correlation between true parameter and estimated items parameter or it can be interpreted that, the greater the N, the stabler the item parameter estimation.

Using the same procedures as those of Problem No. 4, the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication for each magnitude N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in one class in each junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated examinee parameter (Θ) obtained is as follows.

Table XVI. The Mean Correlation between True Parameter and Estimated Examinee Parameter (One-Parameter Logistic Model)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.877122	0.934486	0.962466
N288	0.876503	0.928754	0.955984
N1152	0.874533	0.93011	0.956749

Based on the table above, the correlation vs. sample size (N) and test length (n) can be illustrated in the following figure.

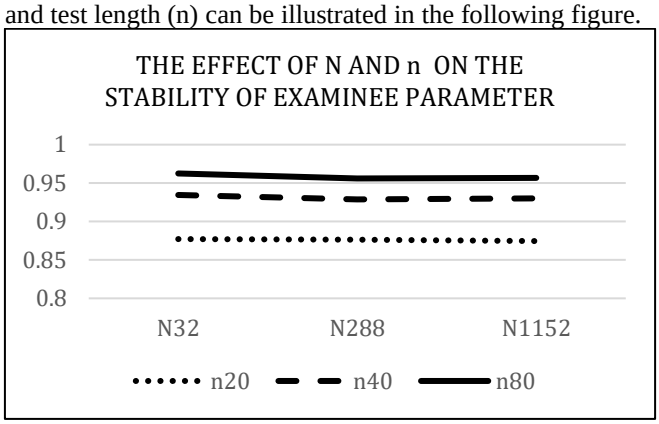

Fig 6. The Effect of N and n on the Stability of Examinee Parameter (One-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that n20 has the lowest correlation line and n80 has the highest correlation line (can be seen from the position of the line that is approaching 1). As for the N value at n80, all three have correlation values that have little different or tend to be stable at the same number. This indicates that the stability of the examinee parameter (Θ) is more influenced by the number of item/test length n than the sample size N. Thus, it can be concluded that the number of n gives a linear effect on the stability of examinee parameter estimation (Θ). The greater the n, the higher the correlation value between true parameter and estimated examinee parameter (Θ) or it can be interpreted that, the bigger the n, the stabler the parameter of examinee (Θ).

2. Two-Parameter Logistic Model

Using the same procedures as those of the previous work, the correlation between true parameter and estimated parameter (can be seen in columns b and a because the model used is two-parameter logistic model) with 5 times replication for each N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in one class in each junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items in Junior High School

National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated item parameter obtained is as follows.

Table XVII. The Average Correlation between True Parameter and Estimated Item b Parameter (Two-Parameter Logistic Model)

AVERAGE CORRELATION (PARAMETER b)			
	n20	n40	n80
N32	0.8344	0.9008	0.7302
N288	0.9882	0.987	0.9862
N1152	0.9988	0.9956	0.9956

Table XVIII. The Average Correlation between True Parameter and Estimated Item a Parameter (Two-Parameter Logistic Model)

AVERAGE CORRELATION (PARAMETER a)			
	n20	n40	n80
N32	0.387	0.5176	0.5666
N288	0.8534	0.9092	0.9132
N1152	0.969	0.9738	0.9782

Based on the tables above, the correlation vs. sample size (N) and test length (n) can be illustrated in the figures below.

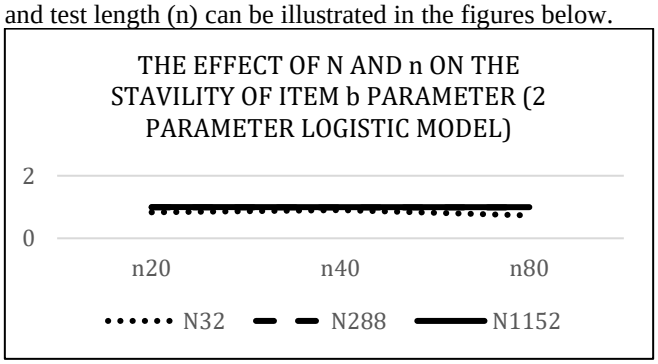

Fig 7. The Effect of N and n on the Stability of Item b Parameter (Two-Parameter Logistic Model)

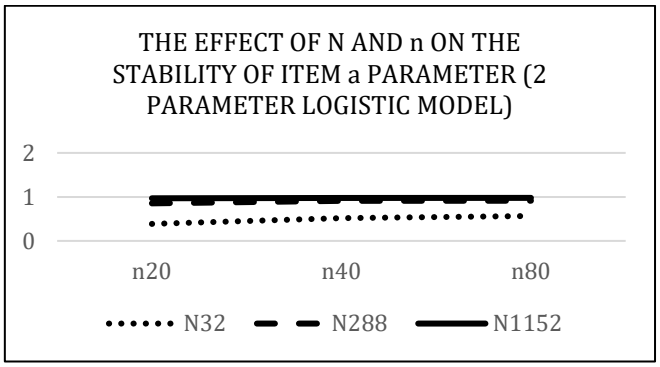

Fig 8. The Effect of N and n on the Stability of Item a Parameter (Two-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graphs of both parameter b and parameter a, it can be interpreted that N32 has the lowest correlation line and N1152 has the highest correlation line (can be seen from the position of the line that is approaching 1). As for the value of n in N1152, all three have correlation values have little different or tend to be stable at the same number. This shows that in two-parameter logistic model, the stability of item parameter is more influenced by sample size N than by the number of item/test length n. In general, it can be concluded that the number of N gives a linear effect on the stability of item parameter estimation. The greater the N, the higher the correlation between true parameter and estimated items parameter or it can be interpreted that, the greater the N, the stabler the item parameter estimation.

Using the same procedures as those of the previous work, the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication, for each number of N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in one class in each junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated examinee parameter (Θ) obtained is as follows.

Table XIX. The Average Correlation between True Parameter and Estimated Examinee Parameter (Two-Parameter Logistic Model)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.923157	0.9328	0.965979
N288	0.914294	0.940279	0.961051
N1152	0.914864	0.942079	0.961316

Based on the table above, the correlation vs. sample size (N) and test length (n) can be illustrated in the figure below.

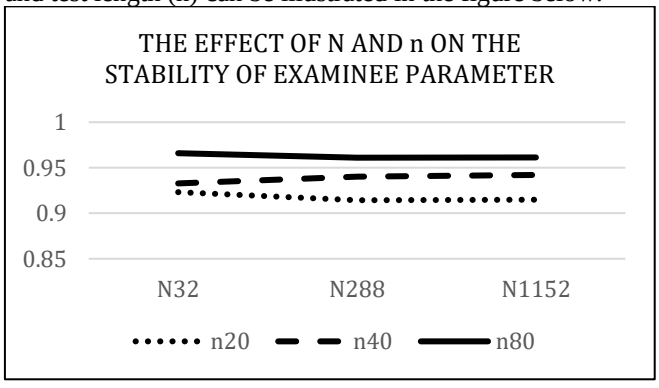

Fig. 9. The Effect of N and n on the Stability of Examinee Parameter (Two-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that n20 has the lowest correlation line and n80 has the highest correlation line (can

be seen from the position of the line that is approaching 1). As for the N value at n80, all three have correlation values have little different or tend to be stable at the same number. This indicates that the stability of examinee parameter (Θ) is more influenced by the number of item/test length n than the sample size N. Thus, it can be concluded that the number of n gives a linear effect on the stability of examinee parameter estimation (Θ). The greater the n, the higher the correlation value between true parameter and estimated examinee parameter (Θ) or it can be interpreted that, the bigger the n, the stabler the examinee parameter (Θ).

3. Three-Parameter Logistic Model

Using the same procedures as those of the previous work, the correlation between true parameter and estimated parameter (can be seen in columns b, a and c because the model used is three-parameter logistic model) with 5 times replication for each N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in one class in each junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated parameter obtained is as follows.

Table XX. The Mean Correlation between True Parameter and Estimated Item b Parameter (Three-Parameter Logistic Model)

AVERAGE CORRELATION (PARAMETER b)			
	n20	n40	n80
N32	0.8904	0.7836	0.6514
N288	0.9816	0.9656	0.9592
N1152	0.99	0.9818	0.978

Table XXI. The Mean Correlation between True Parameter and Estimated Item a Parameter (Three-Parameter Logistic Model)

AVERAGE CORRELATION (PARAMETER a)			
	n20	n40	n80
N32	0.3756	0.5004	0.4052
N288	0.859	0.8242	0.809
N1152	0.8908	0.9168	0.9226

Table XXII. The Average Correlation between True Parameter and Estimated Item c Parameter (Three-Parameter Logistic Model)

AVERAGE CORRELATION (PARAMETER c)			
	n20	n40	n80
N32	0.245333	0.1004	0.1048
N288	0.6102	0.478	0.4882
N1152	0.645	0.6012	0.5882

Based on the tables above, the correlation graph vs. sample size (N) and test length (n) can be summarized in the figures below.

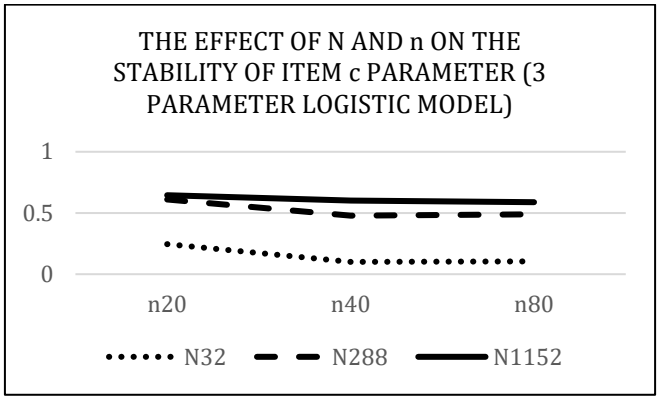

Fig 10. The Effect of N and n on the Stability of Item c Parameter (Three-Parameter Logistic Model)

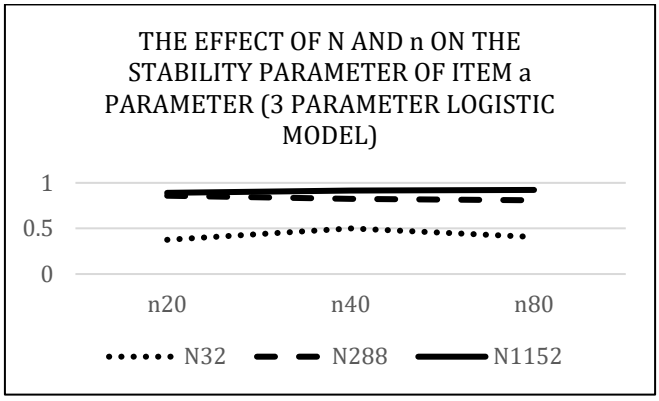

Fig 11. The Effect of N and n on the Stability of Item a Parameter (Three-Parameter Logistic Model)

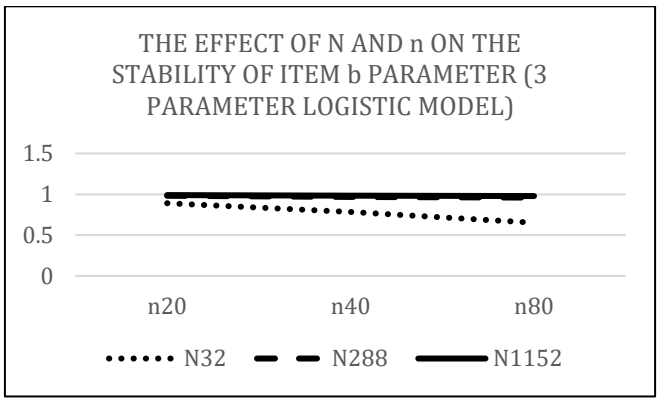

Fig 12. The Effect of N and n on the Stability of Item b Parameter (Three-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graphs of parameter b, parameter a and parameter c above, it can be interpreted that N32 has the lowest correlation line and N1152 has the highest correlation line (can be seen from the position of the line approaching 1). As for the value of n in N1152, all three have correlation values have little different or tend to be stable at the same number. This shows that in three-parameter logistic model, the stability of item parameter is more influenced by sample size N than by number of item/test length n. In general, it can be concluded that the number of N gives a linear effect on the stability of item parameter estimation. The greater the N, the higher the

correlation between true parameter and estimated items parameter or it can be interpreted that, the greater the N, the stabler the item parameter estimation.

Using the same procedures as those of the previous work, the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication for each N (32, 288 and 1152) and n (20, 40 and 80) can be obtained. The number of N is determined based on the assumptions that the maximum number of students in one class in each junior high school is 32, and if one school consists of 9 classes then the number of students is 288 and if one district consists of 4 public schools, the total number of students is 1152. The number of n is determined based on assumptions that the number of items in Junior High School National Exam is 40 items, and researcher wants to identify the difference between the correlations between true parameter and estimated examinee parameter (Θ) if the number of items is shortened to ½ time, that is 20, and if it is lengthened twice, that is 80.

Using N (32, 288 and 1152) and n (20, 40 and 80), the average value of correlation between true parameter and estimated examinee parameter (Θ) obtained is as follows.

Table XXIII. The Average Correlation between True Parameter and Estimated Examinee Parameter (Three-Parameter Logistic Model)

AVERAGE CORRELATION			
	n20	n40	n80
N32	0.832816	0.90596	0.957815
N288	0.889316	0.92327	0.956889
N1152	0.89128	0.922902	0.955814

Based on the table above, the correlation graph vs. sample size (N) and test length (n) can be summarized in the following figure.

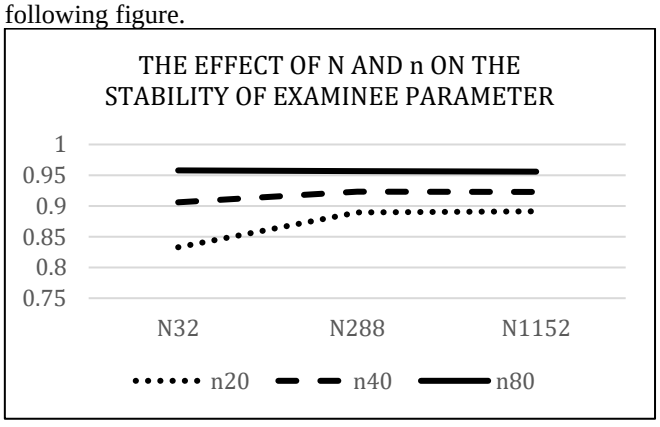

Fig 13. The Effect of N and n on the Stability of Examinee Parameter (Three-Parameter Logistic Model)

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that n20 has the lowest correlation line and n80 has the highest correlation line (can be seen from the position of the line that is approaching 1). As for the N value at n80, all three have correlation values have little different or tend to be stable at the same number. This indicates that the stability of the examinee parameter (Θ) is more influenced by the number of item/test length n than by the sample size N. Thus, it can be concluded that the number of n gives a linear effect on the stability of examinee parameter estimation (Θ). The greater the n, the higher the correlation value between true parameter and estimated examinee parameter (Θ) or it can be interpreted that the bigger the n, the stabler the examinee parameter (Θ).

4. All Logistic Models

To determine the effect of the model on the stability of item parameter estimation and examinee parameter estimation (Θ), it can be seen from comparison of correlation between models.

Using the same procedures as those of the previous work, the correlation between true parameter and estimated item parameter (which is seen in column b because all the models have parameter b, so it can be compared) with 5 times replication, for each model (one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model)

From each model with N1152 and n80, the average value of correlation between true parameter and estimated item parameter item obtained is as follows.

Table XXIV. The Average Correlation between True Parameter and Estimated Item Parameter (All Logistic Models)

	AVERAGE CORRELATION
One-parameter logistic model	0.998
Two-parameter logistic model	0.9956
Three-parameter logistic model	0.978

Based on the table above, the graph correlation vs model can be illustrated in the following figure.

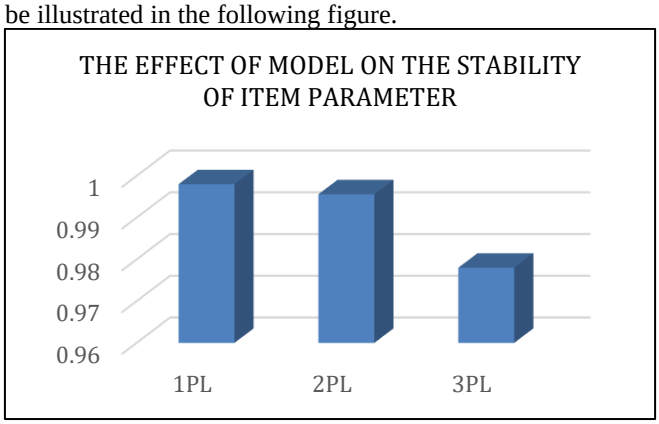

Fig 14. The Effect of Model on the Stability of Item Parameter

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that three-parameter logistic model has the lowest correlation and one-parameter logistic model has the highest correlation (close to 1). In this case, based on the number of the correlation, one-parameter logistic model is the best model in estimating the item parameter, although the difference of of the correlation between one-parameter logistic model and two-parameter logistic model is quite small. The low correlation between true parameter and estimated item parameter in three-parameter logistic model shows that three-parameter logistic model is the most unstable because the correlation value between true parameter and estimated item parameter is the lowest.

Using the same procedures of those of the previous work, the correlation between true parameter and estimated examinee parameter (Θ) with 5 times replication for each model (one-parameter logistic model, two-parameter logistic model, and three-parameter logistic model) can be obtained.

From each model with N1152 and n80, the average value of correlation between true parameter and estimate examinee parameter can be summarized in the following table.

Table XXV. The Average Correlation between True Parameter and Estimated Examinee Parameter (All Logistics Models)

	AVERAGE CORRELATION
One-parameter logistic model	0.956749
Two-parameter logistic model	0.961316
Three-parameter logistic model	0.955814

Based on the table above, the correlation graph vs models can be illustrated in the following figure.

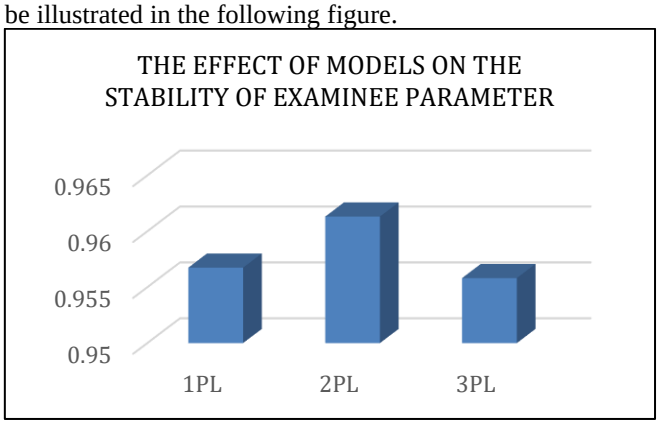

Fig 15. The Effect of Models on the Stability of Examinee Parameter

What is meant by the stability criterion here is the greater the correlation, the stabler the parameter estimation. Based on the graph above, it can be interpreted that if the models are ranked, two-parameter logistic model has the highest correlation (close to 1), then followed by one-parameter logistic model and the last is three-parameter logistic model as it has the lowest correlation. The low correlation between true parameter and estimated examinee parameter in three-parameter logistic model shows that three-parameter logistic model is the most unstable as the correlation value between true parameter and estimated examinee parameter is the lowest. In this case, based on the correlation, the two-parameter logistic model is the best model to estimate the examinee parameter.

CONCLUSION

Based on the results and previous discussion, the following conclusions can be drawn.

First, the bigger the N, the higher the correlation value between true parameter and estimated item parameter, so it can be interpreted that the greater the N, the stabler the item parameter estimation.

Second, the more n, the higher the correlation value between true parameter and estimated examinee parameter (Θ) or it can be interpreted that the more n, the stabler the estimated examinee parameter (Θ).

Third, based on the correlation, (a) to estimate item b parameter, it is best to use the one-parameter logistic model because two-parameter logistic model and three-parameter logistic model are not as stable as one-parameter logistic model in estimating item b parameter; (b) to estimate the item a parameter, it is better to use two-parameter logistic model as the correlation in one-parameter logistic model is all 0, whereas the three-parameter logistic model is not as stable as two-parameter logistic model; (c) to estimate the item c parameter, it is better to use three-parameter logistic model because one-parameter logistic model and two-parameter logistic model have no guessing factor. The guessing factor is equal to 0, so the low-ability person is assumed that c=0; (d) two-parameter logistic model is the most stable model in estimating the examinee parameter.

Fourth, based on the correlation: (a) The stability of the item parameter is more influenced by the sample size N than the number of item/test length n. The greater the N, the higher the correlation value between the true parameter and estimated item parameter; (b) The stability of the examinee parameter (Θ) is more influenced by the number of item/test length n than the sample size N. The greater the n, the higher the value of the correlation between true parameter and estimated examinee parameter (Θ).

Fifth, based on the magnitude of the correlation: (a) In one-parameter logistic model, the stability of the item parameter is more affected by the sample size N than the number of item/test length n. The greater the N, the higher the correlation value between the true parameter and estimated item parameter; (b) In one-parameter logistic model, the stability of examinee parameter (Θ) is more affected by the number of item/test length n than sample size N. The greater the n, the higher the value of the correlation between true parameter and estimated examinee parameter (Θ); (c) In two-parameter logistic model, the stability of the item parameter is more influenced by the sample size N than the number of item/test length n. The greater the N, the higher the correlation value between true parameter and estimated item parameter; (d) In two-parameter logistic model, the stability of examinee parameter (Θ) is more influenced by the number of item/test length n than sample size N. The greater the n, the higher the value of the correlation between true parameter and estimated examinee parameter (Θ); (e) In three-parameter logistic model, the stability of the item parameter is more influenced by the sample size N than the number of item/test length n. The greater the N, the higher the correlation value between true parameter and estimated item parameter; (f) In three-parameter logistic model, the stability of examinee parameter (Θ) is more influenced by number of item/test length n than sample size N. The bigger the n, the higher the value of correlation between true parameter and estimated examinee parameter (Θ); (g) The one-parameter logistic model is the best model to estimate item parameter, although the difference of the correlation between one-parameter logistic model and two-parameter logistic model is quite small and three-parameter logistic model is most unstable because the correlation value between true parameter and estimated item parameter is the lowest; (h) three-parameter logistic model is the most unstable because the correlation value between true parameter and estimated examinee

parameter examinee is the lowest, while two-parameter logistic model is the best model in estimating examinee parameter.

REFERENCES

[1] Akour, M., and AL-Omari, H., 2013, "Empirical Investigation of the Stability of Irt Item-Parameters Estimation", International Online Journal of Educational Sciences, vol. 5, pp.

[2] Allen, M. J., and Yen, W. M., 2001, Introduction to Measurement Theory, Waveland Press,

[3] Chuah, S. C., Drasgow, F., and Luecht, R., 2006, "How Big Is Big Enough? Sample Size Requirements for Cast Item Parameter Estimation", Applied Measurement in Education, vol. 19, pp. 241-55.

[4] Crocker, L., and Algina, J., 1986, Introduction to Classical and Modern Test Theory, ERIC,

[5] Cronbach, L. J., 1990, Essentials of Psychological Testing, happer and Row publishers, New York

[6] Goldman, S. H., and Raju, N. S., 1986, "Recovery of One- and Two-Parameter Logistic Item Parameters: An Empirical Study", Educational and Psychological Measurement, vol. 46, pp. 11-21.

[7] Guyer, R., and Thompson, N., 2012, "User's Manual for Xcalibre Item Response Theory Calibration Software, Version 4.1. 8", St. Paul, MN: Assessment Systems Corporation.[Links], vol., pp.

[8] HAMBLETON, R. K., and COOK, L. L., 1983, Robustness of Item Response Models and Effects of Test Length and Sample Size on the Precision of Ability Estimates, in New Horizons in Testing, Elsevier,

[9] Hambleton, R. K., Swaminathan, H., and Rogers, H. J., 1991, Fundamentals of Item Response Theory, Sage,

[10] Harwell, M. R., and Janosky, J. E., 1991, "An Empirical Study of the Effects of Small Datasets and Varying Prior Variances on Item Parameter Estimation in Bilog", Applied Psychological Measurement, vol. 15, pp. 279-91.

[11] Heri, R., 2006, "Stabilitas Estimasi Parameter Pada Regresi Logistik (Suatu Penerapan Pada Pengukuran)", Trend Penelitian dan Pembelajaran Matematika di Era ICT, vol., pp.

[12] Hulin, C. L., Lissak, R. I., and Drasgow, F., 1982, "Recovery of Two-and Three-Parameter Logistic Item Characteristic Curves: A Monte Carlo Study", Applied psychological measurement, vol. 6, pp. 249-60.

[13] Lim, R. G., and Drasgow, F., 1990, "Evaluation of Two Methods for Estimating Item Response Theory Parameters When Assessing Differential Item Functioning", Journal of Applied Psychology, vol. 75, pp. 164.

[14] Lord, F. M., 1967, "An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Logistic Model", ETS Research Report Series, vol. 1967, pp.

[15] Patsula, L., 1995, A Comparison of Item Parameter Estimates and Iccs Produced with Testgraf and Bilog under Different Test Lengths and Sample Sizes, University of Ottawa (Canada),

[16] Ree, M. J., and Jensen, H. E., 1983, Effects of Sample Size on Linear Equating of Item Characteristic Curve Parameters, in New Horizons in Testing, Elsevier,

[17] Sahin, A., and Anil, D., 2017, "The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory", Educational Sciences: Theory and Practice, vol. 17, pp. 321-35.

[18] Stone, C. A., 1992, "Recovery of Marginal Maximum Likelihood Estimates in the Two-Parameter Logistic Response Model: An Evaluation of Multilog", Applied Psychological Measurement, vol. 16, pp. 1-16.

[19] Suwarto, S., 2005, "Pengaruh Ukuran Sampel Dari Model Logistik Terhadap Estimasi Parameter Item", JURNAL PENDIDIKAN, vol. 14, pp.

[20] Swaminathan, H., and Gifford, J. A., 1983, Estimation of Parameters in the Three-Parameter Latent Trait Model, in New Horizons in Testing, Elsevier,

[21] Tang, K. L., Way, W. D., and Carey, P. A., 1993, "The Effect of Small Calibration Sample Sizes on Toefl Irt-Based Equating", ETS Research Report Series, vol. 1993, pp.

[22] Thissen, D., and Wainer, H., 1982, "Some Standard Errors in Item Response Theory", Psychometrika, vol. 47, pp. 397-412.

[23] Weiss, D. J., and Von Minden, S., 'A Comparison of Item Parameter Estimates from Xcalibre 4.1 and Bilog-Mg', (St. Paul, MN: Assessment Systems Corporation, 2012).

[24] Yen, W. M., 1987, "A Comparison of the Efficiency and Accuracy of Bilog and Logist", Psychometrika, vol. 52, pp. 275-91.

[25] Yoes, M., 1995, "An Updated Comparison of Micro-Computer Based Item Parameter Estimation Procedures Used with the 3-Parameter Irt Model", St. Paul, MN: Assessment Systems Corporation, vol., pp.