Al-Khwarizmi: Jurnal Pendidikan Matematika dan Ilmu Pengetahuan Alam
Maret-2021, Vol.9, No.1, hal.69-78
ISSN(P): 2337-7666; ISSN(E):2541-6499
https://ejournal.iainpalopo.ac.id/index.php/al-khwarizmi
DOI: http://dx.doi.org/10.24256/jpmipa.v9i1.2384

Sample Size and Test Length for Item Parameter Estimate and Exam Parameter Estimate

Riswan
Departemen Statistika, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Pertanian Bogor
Jl. Meranti Wing 22 Level 4, Kampus IPB Darmaga, Bogor, 16680, Jawa Barat
Email: anmriswan@apps.ipb.ac.id

Article History:
Received: 17-11-2020; Received in Revised: 2-03-2021; Accepted: 21-02-2021

Abstract

The Item Response Theory (IRT) model contains one or more parameters in the model. These parameters are unknown, so it is necessary to predict them. This paper aims (1) to determine the sample size (N) on the stability of the item parameter (2) to determine the length (n) test on the stability of the estimate parameter examinee (3) to determine the effect of the model on the stability of the item and the parameter to examine (4) to find out Effect of sample size and test length on item stability and examinee parameter estimates (5) Effect of sample size, test length, and model on item stability and examinee parameter estimates. This paper is a simulation study in which the latent trait (θ) sample simulation is derived from a standard normal population of ~ N (0.1), with a specific Sample Size (N) and test length (n) with the 1PL, 2PL and 3PL models using Wingen. Item analysis was carried out using the classical theory test approach and modern test theory. Item Response Theory and data were analyzed through software R with the ltm package. The results showed that the larger the sample size (N), the more stable the estimated parameter. For the length test, which is the greater the test length (n), the more stable the estimated parameter (θ).

Keywords: *Item Response Theory; Item Stability; Sample Size; Test Lenght; Wingen.*

Abstrak

Model Item Respons Theory (IRT) mengandung satu atau lebih parameter dalam model. Parameter-paramter tersebut tidak diketahui sehingga perlu diduga. Tulisan ini bertujuan (1) untuk mengetahui sample size (N) terhadap kestabilan item parameter (2) untuk mengetahui Test lenght (n) terhadap kestabilan examinee parameter estimate (3) untuk mengetahui pengaruh model terhadap kestabilan item dan parameter examinee (4) untuk mengetahui Pengaruh sampel size dan test length terhadap kestabilan item dan examinee parameter estimates (5) Pengaruh sample size, test length, dan model terhadap kestabilan item dan examinee parameter estimates. Tulisan ini merupakan suatu studi simulasi dimana simulasi sampel latent trait (θ) berasal populasi normal baku ~N(0,1), dengan Size Sample (N) dan test length (n) tertentu dengan model 1PL, 2PL dan 3PL menggunakan Wingen. Analisis item dilakukan dengan pendekatan tes teori klasik dan teori tes modern Item Respons Theory serta data dianalisis melalui softawre R dengan package ltm. Hasil penelitian menunjukkan bahwa semakin besar sample size (N) maka dugaan parameter semakin stabil. Untuk Test lenght yakni semakin besar test length (n) maka dugaan parameter (θ) semakin stabil.

Keywords: *Item Response Theory; Kestabilan Item; Sample Size; Test Lenght; Wingen.*

Introduction

In the development of science, it has given birth to several disciplines, in the fields of education and psychology in studying measurement methods and their solutions has developed into a special discipline called test theory. specific situations, and formulate methods for overcoming and impregnating the problem[1].

Item response theory (IRT) was used intensively in educational measurement in the 1970s and 1990s, to become a statistical tool for modeling multivariate discrete response data. IRT can be used in a wide variety of fields from education, psychology, economics, and demography to medical research[2]. The advantage of IRT is its have strong psychometric properties[3].

Item Response Theory (IRT) is a modern test theory that explains how to make inferences about the characteristics or abilities of test takers that cannot be observed based on participant response data to items. This theory is based on two assumptions, namely (1) the performance of the examinee (examinee performance) on a test item that can be predicted (or can be explained) through a set of factors called ability (trait, latent trait, or ability), and (2) The relationship between the performance of the examinees item and the set of abilities based on the performance of the item and can be described by a monotonous upward function called the item characteristic function or item characteristic curve (ICC). In addition, according to Stark, the item difficulty must also be considered in order to make a test with the desired total score distribution[4].

There are several item response models in which all IRT models contain one or more item parameters and one or more examinee parameters. This paper focuses on the item response model with one examinee parameter. These parameters are not known, so they need to be guessed. Estimation is done by simulation of the data generated using the WinGen program for 5 replications with a certain sample size and test length. Furthermore, the simulation data are analyzed using R software where the analysis stage is based on the correlation and RMSD criteria to see the stability of the estimated parameters.

---

[1] Linda Crocker and James Algina, *Introduction to Classical and Modern Test Theory* (Rinchart and Winston: Inc. Amerika, 1986).

[2] Brian W. Junker, "Factor Analysis and Latent Structure: IRT and Rasch Models," in *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, ed. James D. Wright (Oxford: Elsevier, 2015), 698–702, https://doi.org/10.1016/B978-0-08-097086-8.42126-4.

[3] Everett L. Worthington et al., "Chapter 17 - Measures of Forgiveness: Self-Report, Physiological, Chemical, and Behavioral Indicators," in *Measures of Personality and Social Psychological Constructs*, ed. Gregory J. Boyle, Donald H. Saklofske, and Gerald Matthews (San Diego: Academic Press, 2015), 474–502, https://doi.org/10.1016/B978-0-12-386915-9.00017-6.

[4] S Stark et al., *IRT Modeling Lab: Test Development Using Classical Test Theory* (Urbana: University of Illinois, 2001).

Method

In a simulation study, the simulation of latent trait (θ)samples is derived from the standard normal population of ~ N (0.1), with a specific Sample Size (N) and test length (n) with the 1PL, 2PL and 3PL models using Wingen[5]. Item analysis was carried out using the Classical Test Theory (CTT) approach and the modern Item Response Theory (IRT) theory test, then the data were analyzed through software R with the lTM package.

Result and Discussion

1. Effect of sample size on the stability of item parameter estimates

Referring to the opinion put forward by Nunnally et. al, to see the effect of the sample size on the stability of the item parameters estimates, the number of items is determined by referring to the rule of thumb (test length) n = 40 and the sample size is varied, namely N = 200, 400, 1000[6]. Each of these variations was replicated 5 times using the 1PL model and presented in the correlation and RMSD summary table as follows.

Tabel 1. Correlation and RMSD Parameters

Replication	Correlation			RMSD Parameter b		
	200	400	1000	200	400	1000
1	0.991	0.9960	0.9992	0.354	0.10437	0.04828
2	0.992	0.9950	0.9992	0.422	0.11152	0.06265
3	0.991	0.9979	0.9993	0.282	0.13807	0.08213
4	0.993	0.996	0.9986	0.212	0.11625	0.0636
5	0.993	0.9961	0.999	0.314	0.13640	0.07398

Source: Processed data with the Wingen Application

Tabel 2. Correlation and RMSD Parameters

Sample (N)	b	RMSD
200	0.9924202	0.3172997
400	0.9962736	0.1213278
1000	0.9991467	0.06614813

Source: Processed data with the Wingen Application

[5] Robert L. Linn et al., "Item Bias in a Test of Reading Comprehension," *Applied Psychological Measurement* 5, no. 2 (April 27, 1981): 159–73, https://doi.org/10.1177/014662168100500202.

[6] Jum C. Nunnally and Ira H. Bernstein, *Psychometric Theory*, 3rd ed. (New York: McGraw Hill, 1994).

Based on the summary table above, the following sample size plot is formed with the correlation values as follows.

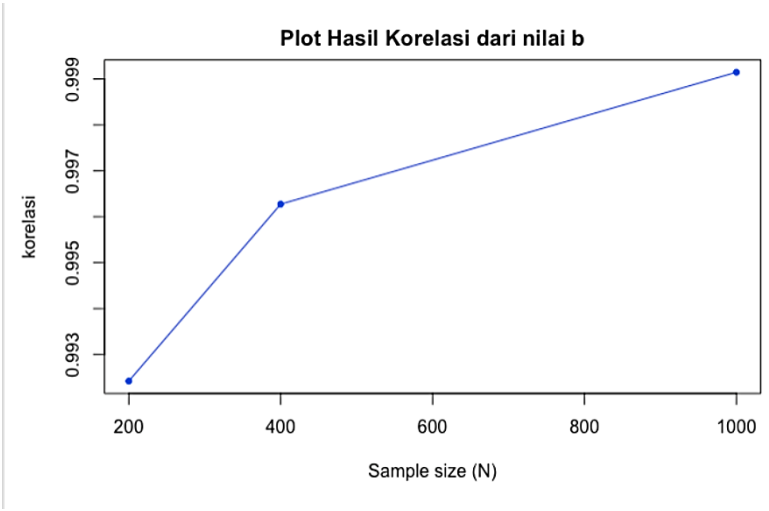


Figure 1. Graph of the correlation results from the value of b

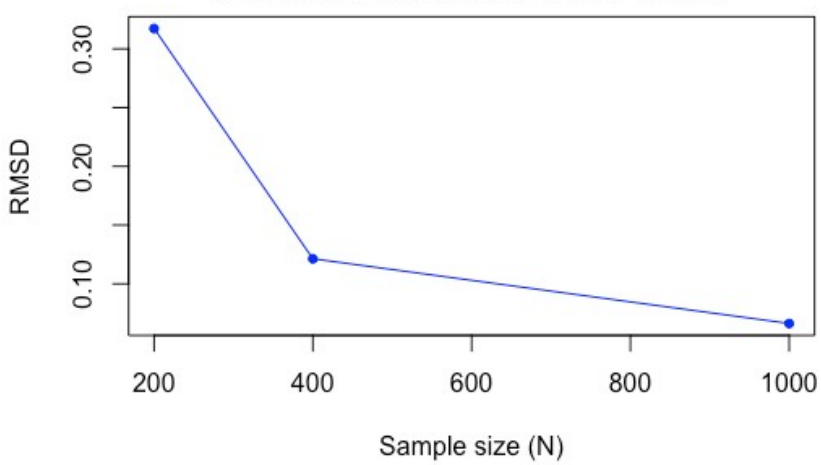


Figure 2. Graph of RMSD correlation results

Based on Figure 1 and Figure 2 above, it shows that the larger the sample size (N), the greater the correlation value, while the RMSD value is getting smaller. So the Sample size (N) affects the stability of the estimated parameters, namely the larger the sample size (N), the more stable the estimated parameter.

2. Effect of test length on the stability of the examinee parameter estimates

To determine the effect of the test length (n) on the stability of the examinee parameter (θ) in the two-parameter logistic model (2PL), the number of examinees (sample size) N = 1000 was determined and the test length was varied, namely n = 20, 40, 100, each variation was replicated 5 times. To see the stability of the estimated examinee (θ) parameter, a correlation coefficient and RMSD (root mean squared differences) or RMSE (root meat squared error) were used. The results of the correlation and RMSD of the examinee parameters can be seen as follows

Table 3. Results of correlation and RMSD parameter examinee(θ)

Test Length	Parameter *Examinee*	
	Correlation	RMSD
20	0.9400224	0.3354105
40	0.9726023	0.2367724
100	0.9879318	0.2553879

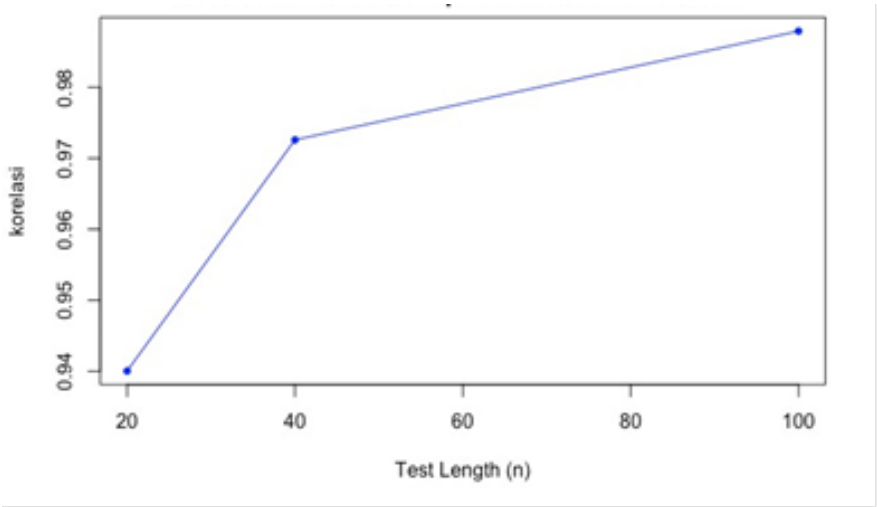


Figure 3. Correlation graph of the examinee parameters

From the correlation chart of the examinee parameters above, it can be concluded that the test length (n) with the sample size (N) will affect the value (θ) true, where the greater the test length, the higher the estimated ability of the sample size. The same thing was also expressed by Xing and Hambleton, the longer the test length, the higher the reliability[7].

[7] Ronald K Hambleton, Hariharan Swaminathan, and H Jane Rogers, *Fundamentals of Item Response Theory* (California: Sage Publications, 1991).

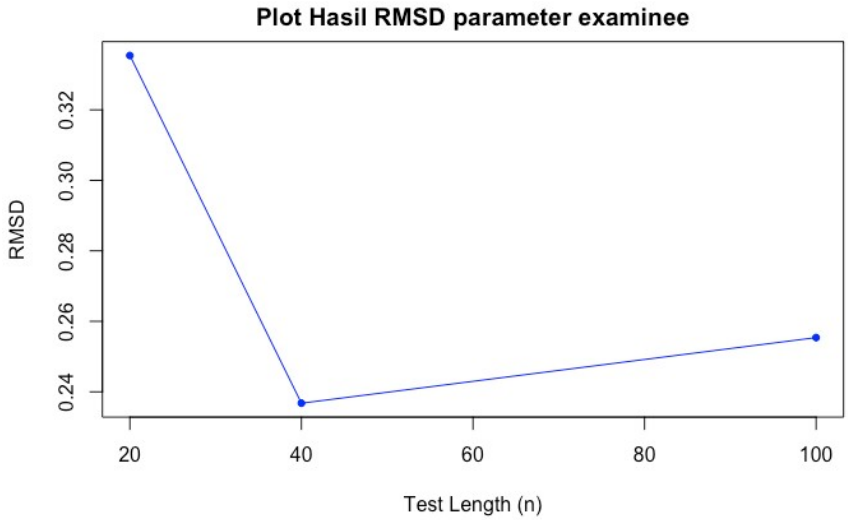


Figure 3. RMSD graph of the examinee parameters

Table 2 with Figure 3 shows that the test length (n) affects the value (θ) true. We see that the RMSD value is smaller when the test length is large, namely the test length is 40, even though when the test length becomes 100, the increase in the RMSD value is not too large. So that the test length (n) affects the stability of the examinee parameter (θ), namely the greater the test length (n), the more stable the estimated parameter (θ).

3. The Effect of the model on the stability of the items and the examinee parameters

With the simulation carried out with the WinGen program with 5 replications with each sample size (N) 400 and test length (n) 40 items by generating WinGen data with 3 models each (1PL, 2PL and 3PL), the correlation results are obtained. on each model the estimanee parameters are below:

Table 4. b and RMSD correlation with the model

Parameter	Kolerasi b	RMSD
1PL	0.9978277	0.1064805
2PL	0.9759783	0.2912042
3PL	0.9814975	0.3604854

Source: Processed data

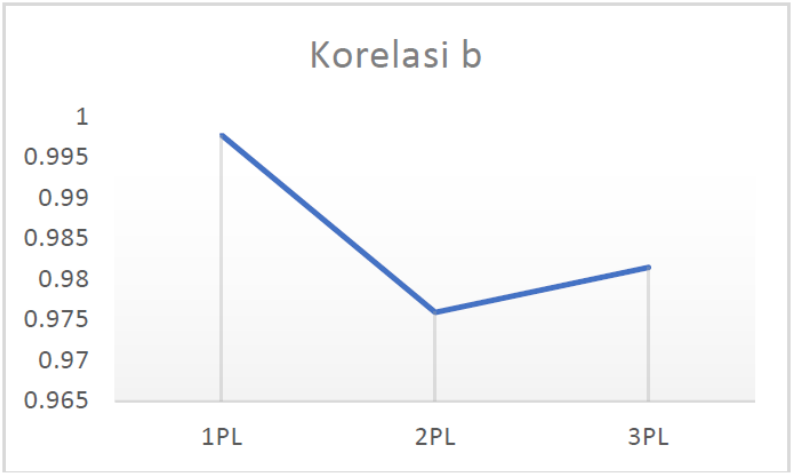


Figure 5. Graph of the results of the correlation b with the model

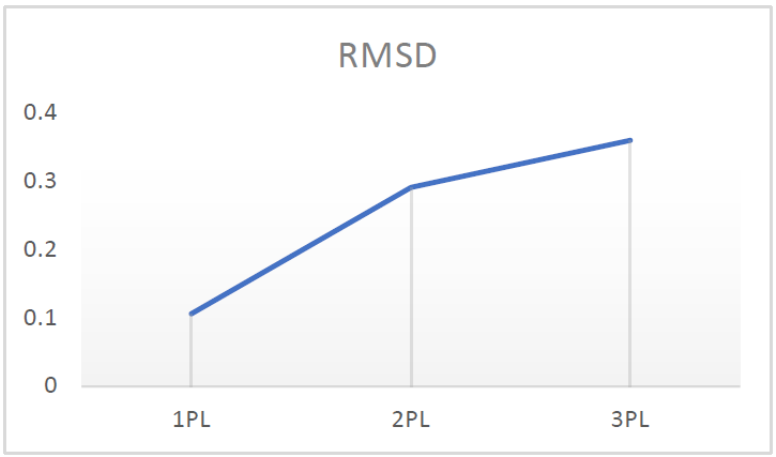


Figure 6. Graph of the results of the correlation b with the model

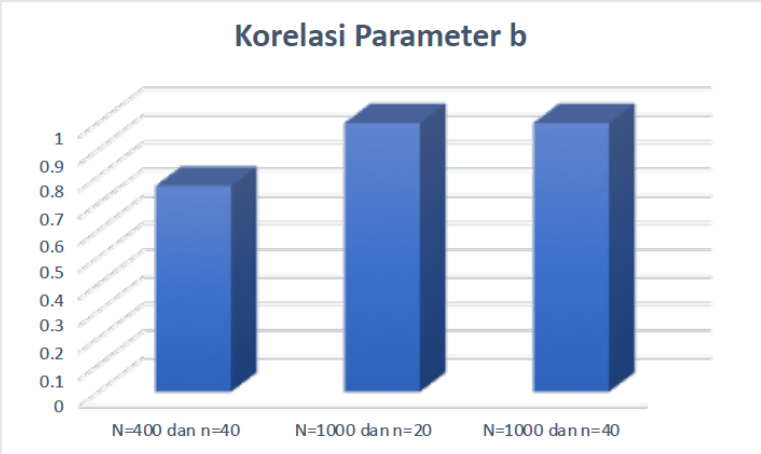
Based on the graph of the RSMD results with the model, it can be seen that the RMSD value will be the opposite of the b parameter value in the model will be higher if more models are used, namely 3 PL.

4. Effect of sample size and test length on item stability and examinee parameter estimates

By doing various simulations, namely by increasing the number of sample size (N) and test length (n) with the 1 PL model then the data is generated through wingen with normal distribution (0.1).

Table 5. Result of Parameter Correlation b

N and n	b
N=400 dan n=40	0.764721
N=1000 dan n=20	0.9980098
N=1000 dan n=40	0.9982648



Figur 7. Graph of correlation result b with model

From the graph above, it can be seen that a large sample size (N) will add to the estimation of parameter b, while the test length (n) which has a lot of influence but is not large on the estimation of the parameter item b.

5. Effect of sample size, test length, and model on item stability and examine parameter estimates

To determine the effect of the number of N and n and the selection of the most appropriate model to get a high b value, a simulation was carried out based on these three aspects. The following is the calibration result of the WinGen output after replicating it five times for each model.

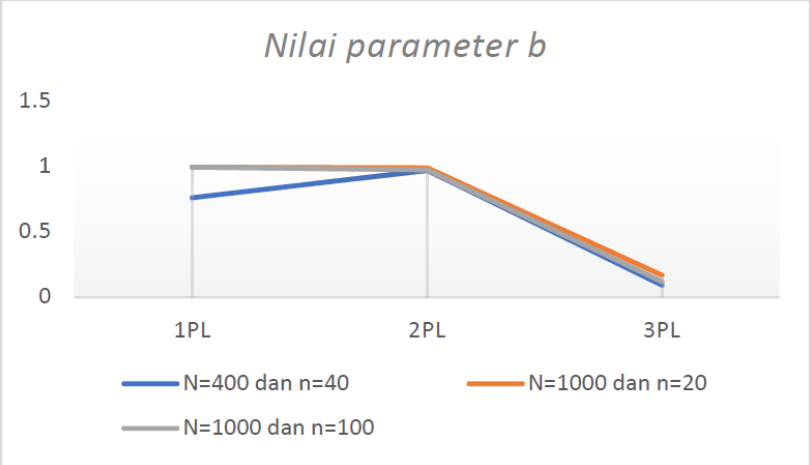


Figure 8 Graph of parameter recapitulation results b

From the graph above shows that the PL model 1 shows the effect of the number of sample size (N) and the test length on the value of b. It can be seen that if N is low and n is high, the correlation value is small, whereas if N is large and n is small, it results in a large correlation. In the 2PL model, the sample size (N) and the large test length (n) produce a high correlation value. But in the 3PL model it can be seen that the number of large N and small n will get a large b value.

Conclusion

The stability of the estimated item parameter is influenced by the sample size, and the stability of the examinee parameter (θ) is influenced by the size of the test length. The larger the sample size, the more stable the item parameter estimate is while the greater the test length, the higher the estimated ability of the sample size. So to maximize the estimated value (θ) it is necessary to look at the sample size, length test, and model used so that if you want to compile a good assessment instrument, you can refer to the sample size, length test and the 1PL, 2PL, and 3PL models.

References

Crocker, Linda, and James Algina. *Introduction to Classical and Modern Test Theory*. Rinchart and Winston: Inc. Amerika, 1986.

Hambleton, Ronald K, Hariharan Swaminathan, and H Jane Rogers. *Fundamentals of Item Response Theory*. California: Sage Publications, 1991.

Junker, Brian W. "Factor Analysis and Latent Structure: IRT and Rasch Models." In *International Encyclopedia of the Social & Behavioral Sciences (Second Edition)*, edited by James D. Wright, 698–702. Oxford: Elsevier, 2015. https://doi.org/10.1016/B978-0-08-097086-8.42126-4.

Linn, Robert L., Michael V. Levine, C. Nicholas Hastings, and James L. Wardrop. "Item Bias in a Test of Reading Comprehension." *Applied Psychological Measurement* 5, no. 2 (April 27, 1981): 159–73. https://doi.org/10.1177/014662168100500202.

Nunnally, Jum C., and Ira H. Bernstein. *Psychometric Theory*. 3rd ed. New York: McGraw Hill, 1994.

Stark, S, S Chernyshenko, D Chuah, Wayne Lee, and P Wilington. *IRT Modeling Lab: Test Development Using Classical Test Theory*. Urbana: University of Illinois, 2001.

Worthington, Everett L., Caroline Lavelock, Charlotte vanOyen Witvliet, Mark S. Rye, Jo-Ann Tsang, and Loren Toussaint. "Chapter 17 - Measures of Forgiveness: Self-Report, Physiological, Chemical, and Behavioral Indicators." In *Measures of Personality and Social Psychological Constructs*, edited by Gregory J. Boyle, Donald H. Saklofske, and Gerald Matthews, 474–502. San Diego: Academic Press, 2015. https://doi.org/10.1016/B978-0-12-386915-9.00017-6.