The Effects of Sample Size and Logistic Models on Item Parameter Estimation

Suwarto[1], Eko Putro Widoyoko[2], Budi Setiawan[3]
suwartowarto@univetbantara.ac.id[1], ekoputro@umpwr.ac.id[2,] , budi.setiawan@umpwr.ac.id

Veteran Bangun Nusantara University of Sukoharjo, Jl. Letjend Sujono Humardani 1 Sukoharjo, Indonesia[1],
Universitas Muhammadiyah Purworejo, Jl. K.H. Ahmad Dahlan 3 Purworejo, Indonesia[23]

Abstract. This research aims to ascertain: (1) the effects of sample size (N) to determine item parameters; (2) the effects of logistic models (1PL, 2PL, and 3PL) on item parameter estimation. The research was conducted with a computer simulation, which was performed in the following steps: by using sample sizes of N=200, N=500, N=1000, each with ten replications. In order to generate the dichotomous data, the DGEN program was used and then these data were performed in the BILOG 3 program. From the DGEN program the "true" item parameters could be ascertained, and from the BILOG 3 program, the estimated item parameters could be ascertained. The criteria used to determine the stability of item parameters were: (1) Seeking the correlation between "true" parameter and the average item parameters to find the highest correlation; and (2) Seeking the MSD average with the smallest variance in each item parameter, which was the best one. The research results show that (1) the order of the effects of sample size from the best to the worst in estimating item parameters were: N=1000, N=500; N=200, (2) the order of logistic models in item parameter estimation were as follows: (a) the order of the logistic models to estimate the difficulty parameter were: IPL (the best), 2PL (medium), 3PL (low); (b) the order of the logistic models to estimate the discrimination parameter were 2PL and then 3PL; (c) the logistic model to estimate the "guessing parameter" was 3PL.

Keywords: sample size, logistic models, item parameters

1 Introduction

In the real world, various phenomena cannot be measured directly but they are measurable through a series of visible indicators and then estimated as a measurement. For example: intelligence, competence, loyalty, proficiency, and so forth. In order to estimate an unseen ability such as intelligence, loyalty, proficiency, and so on, an indicator of an observable behaviour or competency of a test participant is needed. These indicators are arranged to become an instrument that is used to gather the responses of a test participant. By using these responses, the latent competence can be estimated.

In an assessment made in education, a student answers an item in a multiple choice test. Usually the correct answer is awarded a score of 1 and the wrong answer a score of 0. In scoring with the classical test theory (CTT) approach, a student's competency is stated by the total score gained. This procedure does not consider the interaction between a participant and an item. An alternative scoring model that can be used is the item response theory (IRT). The IRT approach is an alternative approach that can be adopted to analyse a test. There are two principles used in this approach, i.e. the relativity principle and the probability principle [1].

ICE 2019, September 28-30, Purwokerto, Indonesia

DOI 10.4108/eai.28-9-2019.2291082

IRT was developed by measurement experts in the field of psychology and education as an effort to minimize the weaknesses in CTT. There are two basic postulates of modern test theory [2], namely: (1) the work of test participant on an item can be predicted from the kind of factors called characteristic, latent characteristic, or ability; (2) relationship between a test participant’s work on an item and the underlying characteristic can be described by the monotonic increasing function called item characteristic function or item characteristic curve (ICC). This function explains that if the level of characteristic (ability) increases, the probability of answering an item correctly also increases. Assumptions in IRT: The first assumption is that in every set of test, the suitability of model to the data can be assessed; this in called overall fit. This second assumption is that in every set of tests, a test item only measures one competency [3]. In other words, when a person is able to answer a difficult item correctly, he/she will certainly be able to answer an easy item [4]. This assumption is called unidimensionality. [5] High ability test takers will has a greater probability of answering correctly when compared with participants who have low ability. The third assumption: the characteristic function of a particular item reflects the real relationship with the ability to respond to an item correctly. The models in IRT are one parameter logistic (1PL), two parameter logistic (2PL) and three parameter logistic (3PL). The main parameter that serves as a basis for calculation in IRT is the participant’s ability. This parameter of the participant’s ability is called θ (=theta). According to Hambleton, the limit of score θ is infinite, and therefore can extend from $-\infty$ to $+\infty$. However, the theta score can be determined in a standard limit from -4 to +4 [6]; [2]; and [7]. in the logistic model in IRT which uses one parameter logistic, an item’s level of difficulty is defined as the value of the participant’s scale of abilities with the probability of 0.50 to answer a particular item correctly [2]. If the value of the difficulty parameter approaches -2, then the difficulty parameter is very low, whereas if the value of the difficulty parameter approaches +2 then the difficulty parameter is very high for a group test taker [8]. In the one parameter logistic model, the probability of a participant’s ability can be made into a mathematical equation as follows.

$$Pi(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \tag{1}$$

The two logistic model parameter uses the parameter of an item’s level of difficulty and an item’s discrimination. The item’s discrimination parameter functions to determine whether an item distinguishes the groups in the measured aspect according to the difference that exists in the groups. By adding a constant, discrimination (a_i) of the item’s discrimination parameter acts as a direction to the normal ogive curve (curve of characteristic). The form of mathematical equation of the two parameter logistic, the probability of a participant’s ability can be made into a mathematical form [2].

$$P_i(\theta) = \frac{e^{D_i a_i (\theta - b_i)}}{1 + e^{D_i a_i (\theta - b_i)}} \tag{2}$$

The three parameter logistics use the parameter of the difficulty level of item b_i, the discrimination level of a_i, and the guessing factor of c_i. The guessing factor is answering correctly by pure chance. This parameter is similar to the probability of answering correctly, so that if the guess is correct, the answer is correct, and if the guess is wrong, the answer is also wrong. If 0 is used for a wrong answer and 1 is used for a correct answer, the probability of a correct answer is in the range of 0 to 1. A items are said to be good if the guessing parameter value is not more than 1/k, with k many choices [9]. The mathematical equation model for 3PL [2] is as follows:

$$P_i(\theta) = c_i + \left\{ (1 - c_i) \frac{e^{D_i a_i (\theta - b_i)}}{1 + e^{D_i a_i (\theta - b_i)}} \right\} \quad (3)$$

According to Hambleton, [2] to make the estimation of parameter stable, a large sample size must be used. To estimate abilities, MLE (Maximum Likelihood Estimation) is used.

It is believed that computer will be utilized in education [10]. This is corroborated by [11] who says that with the development of the computation technology, IRT has developed rapidly. Computer is very much used for a formative test, a summative test, and a diagnostic test. It is also used for the entrance test to a tertiary educational institution, which is a measuring instrument called Computer Based Test (CBT). According to [12], CBT is used for a short test (11-20 items) and a long test (more than 20 items).

The result of research by [13]: (1) The IPL model, the average of the standard errors (SE) and the root mean square error (RMSE) decrease when the sample size increases, and the value of parameter b is not sensitive to the number of items in a test; (2) The 2PL model, the average SE decreases for parameters a and b when the sample size increases, and the RMSE decreases because the sample size and number of items increase, and (3) the 3PL model, the average SE and RMSE decrease when the sample size increases for parameters a and b; the SE of parameter c slightly increases when the sample size increases.

Results of research by [14]: (1) A sample size of 150 (N=150) is acceptable for the IPL model for all test lengths (n=10, n=20, n=30); (2) 2PL model, minimum N=750 for n=10, minimum N=500 for n=20, and minimum N=250 for n=30; (3) 3PL model, minimum N=750 for n=10, minimum N=750 for n=20, and minimum N=350 for n=30). Results of research by [15]: (1) The item's difficulty parameter is the most sensitive to the differences of sample size and abilities especially theta ability under 0; (2) The item's Difficulty parameter is the most sensitive to sample size and test length; and (3) the minimum requirement of N=500 for n=30 for an accurate item parameter estimation, but N=200 for item parameter estimation is still acceptable for n>15.

In connection with the above research results, the present writer carried out a research with n=20 and N=200, N=500, and N=1000. The research used n=20 and N=200 for the reason that N=200 for item parameter estimation is still acceptable for n>15 [15] and N=1000 for the reason that a minimum sample of 1000 is generally recommended for estimating the 3PL model [16]; [17]; and [18]. This is corroborated by [14] that N=750 can already be used for determining item parameters accurately in the 3 logistic models. The research used short tests for the reason that a short test is often used for selection of new students in both state and private universities. This research was conducted with a computer simulation for the reason that it could perform 10 replications in each occurrence so that the results would be more convincing. This research aimed to ascertain (1) the effects of sample size to determine item parameters, and (2) the effects of the logistic models (1PL, 2PL, 3PL) on item parameter estimation.

2 Method

The research was conducted by performing a computer simulation. The simulation was done with the following procedure: using a sample of N=200, N=500, N=1000, each with 10 replications. After the control file was made and done with DGEN program it would produce four files, each with extensions OUT, DAT, ITM, ABL. Then the file with DAT extension was performed with BILOG 3 program (previously the control file for BILOG had been made). BILOG went through three phases so that three files were produced (the file phase), each with extensions PH1, PH2, PH3, and one file with extension PAR (so that from BILOG four files were produced). After each sample size, was replicated ten times, 240 files were produced.

Results of this simulation were compared between the true item parameter (result from the DGEN program) and item parameter estimation (result of the BILOG 3 program) by using the correlation method and the MSD (Mean Squared Deviation) method, in which calculation was done with the excel program. MSD=variance+bias. The criteria used to determine the stability of item parameter were: (1) seeking the correlation between the "true" parameter and the average item parameter to find the most highly correlated, and (2) seeking the MSD average with the smallest variance in each item parameter, which was the best.

3 Results and Discussion

The results of calculation with excel obtained are as follows.

Table 1. Summary of the Effects of Sample size on Item Parameter Estimation with the Correlation Method

Parameter	N	R	Ranking
b	200	0.992857	2
	500	0.99091	3
	1000	0.995569	1
a	200	0.889461	3
	500	0.964262	2
	1000	0.982384	1
c	200	0.555324	3
	500	0.688567	2
	1000	0.837745	1

Note. a= the discrimination parameter; b= the difficulty parameter; c= the guessing parameter; N=sample size; r=pearson correlation

From Table 2 it can be deduced that: in order to estimate b (difficulty level of items), (the sample size for) parameter estimation from the best to the worst are (1) Sample size of 1000, (2) Sample size of 200, and (3) Sample size of 500. In order to estimate a, the sample size, for the estimation of parameter with the sequence from the best to the worst were (1) a sample of 1000, (2) a sample of 500, and (3) a sample of 200. To estimate c (item's guessing parameter), sample size for parameter estimation with the sequence from the best to the worst were: (1) a sample of 1000, (2) a sample of 500, and (3) a sample of 200.

The correlation method to estimate a and c: the result conforms with the theory that the larger the N the better the parameter estimation [2] so that the sequence of the best to the worst were N=1000, N=500, N=200. However, with the correlation method the results were not very satisfactory because the sequence to estimate b was N=1000, N=200, N=500. This was not consistent with the theory [2]. Therefore the investigation preceded to the MSD method.

Result of calculation with the MSD method

Calculation with the MSD method was done with the aid of the excel program. Summary of the estimation of b with the MSD method can be seen in Table 2.

Table 2. Summary of the Estimation of b with the MSD method

Sample Size (N)	MSD average	Bias average	Variance average
1000	0.025389	0.015911	0.009478
500	0.043931	0.029996	0.013935
200	0.065184	0.03554	0.029644

As to the results of the effects of sample size to determine of b, from the best to the worst were as follows: (1)N=1000 with MSD average=0.025389, (2) N=500 with MSD average=0.043931, and (3) N=200 with MSD average=0.065184.

Summary of estimation of a with the MSD method can be seen in **Table 3**.

Table 3. Summary of Estimation of a with the MSD method

Sample Size (N)	MSD average	Bias average	Variance average
1000	0.049648	0.015969	0.03368
500	0.079965	0.030434	0.049532
200	0.124351	0.05779	0.06656

As to the results of the effects of sample size to determine of a, from the best to the worst were as follows: (1)N=1000 with MSD average=0.049648, (2) N=500 with MSD average=0.079965, and (3) N=200 with MSD average=0.124351.Summary of estimation of c with the MSD method can be seen in Table 4.

Table 4. Summary of Estimation of c with MSD method

Sample Size (N)	MSD average	Bias average	Variance Average
1000	0.003326	0.002605	0.000721
500	0.004934	0.004153	0.000781
200	0.005494	0.004782	0.000712

As to the results of the effects of sample size to determine of c, from the best to the worst were as follows: (1)N=1000 with MSD average=0.003326, (2) N=500 with MSD average=0.004934, and (3) N=200 with MSD average=0.005494.

Effects of Logistic Models (1PL, 2PL, and 3PL)on Estimation of b

The results of the effects of logistic models on b estimation, from the best to the worst were as follows:

Table 5. Summary of Analysis of the Effects of Logistic Models on b Estimation

Model	MSD average (N=1000)	MSD average (N=500)	MSD average (N=200)
1PL	0.003307	0.010252	0.023222
2PL	0.007776	0.012721	0.282872
3PL	0.025389	0.043931	0.065184

From Table 5, on the whole it can be deduced that the effects of logistic models on estimation of b, the best was1PL model followed 2PL model and then 3PL model. For N=200, the results were not consistent with the theory because N=200 was an inadequate sample size so that the results were not accurate. In general, for and N=500 and N=1000, the best sequence of the best models was constant: 1PL model, 2PL model, and then 3PL model.

Effects of Logistic Models (2PL, and 3PL) on a Estimation

The results of the effects of logistic models (2PL and 3PL) on estimation of a from the best to the worst were as follows.

Table 6. Summary of Analysis of the Effects of Logistic Models on a Estimation

Model	MSD average (N=1000)	MSD average (N=500)	MSD average (N=200)
2PL	0.013939	0.022645	0.045805
3PL	0.049648	0.079965	0.1243351

From Table 6, in general it can be deduced that the effects of logistic models (2PL and 3PL) on estimation of a, were as follows: The best was 2PL followed by 3PL.

Effects of Logistic Model (3 PL) on c Estimation

The effects of logistic models on the estimation of c only occurs with one model, i.e. 3PL. This is because there is no factor c or c=0 for 1PL and 2PL.

The results of calculation with MSD show that there are effects of sample size on the estimation of parameter b from the best to the worst, i.e. (1) N=1000 with MSD average=0.025389; (2) N=500with MSD average=0.043931; (3) N=200 with MSD average=0.065184. Similarly, it was found by [13] that SE and RMSE averages decrease when the sample size increases. [15] stated that b is the most sensitive to different sample sizes.

The effects of sample size to determine of parameter a from the best to the worst are: (1) N=1000 with MSD average=0.049648; (2) N=500 with MSD average=0.079965; (3) N=200 with MSD average=0.124351. [13] stated that the SE average decreases for parameters a and b when the sample size increases in 2PL model, and the SE and RMSE averages decrease when the sample size increases for parameters a and b in 3PL model.

The effects of sample size to determine of c parameter, from the best to the worst are: (1) N=1000 with MSD average=0.003326; (2) N=500 with MSD average=0.004934; (3) N=200 with MSD average=0.005494. [12] stated that SE average for the estimation of c parameter slightly increases when the sample size increases in 3PL model. In general, the findings are consistent with a statement in the book *Fundamentals of Item Response Theory*, written by [2] and the results of research [15]; [14]; and [13]. The research performed a computer simulation with various sample sizes using a test with a constant number of items, i.e. 20 items by [12] regarded as a short test. This was also done by [19] in a simulation done with different sample sizes, distribution of abilities, and a test with the same number of items.

This research of the effects of sample size to determine item parameter is in accord with the proposal by [20] to carry out a further study of different sample sizes to determine item parameters.

The effects of logistic models (1PL, 2PL and 3PL) to determine of b parameter: the best is1PL model followed by 2PL model and 3PL model. Reasons: (1) 1PL model needs the least prerequisites, i.e only b parameter, whereas a and c parameters are made constant; (2) 2PL model needs b and a parameters, whereas c parameter is made constant; (3) 3PL model needs all item parameters: b,a, and c parameters are prerequisites, so that the fewest the prerequisites needed, the more accurate the estimation of item parameters.

The effects of 2PL logistic models (2PL and 3PL) on the estimation of parameter a: the best is of 2PL model followed by 3PL model. Reasons: (1) 2PL model needs the least prerequisites, i.e. b and a parameters, whereas c parameter is made constant; (2) 3PL model needs all item parameters: b, a, and c parameters are prerequisites so that the fewest the prerequisites needed, the more accurate the estimation of item parameters. Estimation of parameter c can only be done with 3PL model because 1 PL model and 2PL model do not need parameter c (c=0). Therefore estimation of parameter c can only be done with 3PL model.

The effects of the three logistic models on the estimation of b parameter: it can be deduced that the best effect is1PLmodel followed by 2PL model and 3PL model. For N=200, the results are not consistent with the theory because N=200 is an inadequate sample size so that the results are not accurate. This is congruent with the findings by [13]. In general, for N=500 and N=1000, the sequence/order of the best model remains constant: 1PL model, 2PL model, and3PL model. The best effect of logistic models (2PL and 3PL) on estimation of parameter a is 2PL model followed by 3PL because for 1PL model there is no factor a. The effect of logistic models (3PL) on parameter c only occurs in one model (i.e. 3PL) because in 1PL model and 2PL model there is no guessing factor or c=0.

4 Conclusion

The sequence or order of sample sizes from the best to the worst on estimating item parameters (a, b, c parameters) in the simulation is as follows: (1) Sample sizes of N=1000, N=500, and N=200. For the try-out and data collection using an instrument (test) in which in calculating the results a modern test theory is used (i.e. IRT) a big sample size is needed. The sequence of logistic models (1PL, 2PL, and 3PL) to determine item parameter is as follows: the sequence of models to estimate parameter b: (1) 1PL model is the best, (2) 2PL model is medium, and (3) 3PL model is low. The sequence of models to estimate parameter a: (1) 2PL model is the best followed by (2) 3PL model. The logistic model to estimate parameter c is 3PL model. Estimation of parameter c only occurs in 3PL model.

References

[1] Keeves, J.P., & Alagumalai, S. New appoaches to measurement. In Masters, G.N. dan Keeves, J.P. (Eds). Advances in measurement in educational research and assesment. Amsterdam, Pergamon. (1999).

[2] Hambleton, R.K., Swaminathan, H., & Rogers, H.J. Fundamental of item response theory. Newbury Park, CA, Sage. (1991).

[3] Suryabrata, S. Pengembangan alat ukur psikologis. Andi, Yogyakarta. (2000).

[4] Mardapi, D. Peran statistika pada bidang pengukuran pendidikan. Makalah disampaikan pada Seminar Peran Statistik dalam Bidang ilmu Pengetahuan, di FMIPA UGM. (2000).

[5] Suwarto. Discrimination, difficulty, and guessing the biology test 8th grade by the period of the odd term. Proceeding Biology Education Conference (ISSN: 2528-5742), Vol 13 (1) 2016: 151-158. https://jurnal.uns.ac.id/prosbi/article/viewFile/5680/5048

[6] Naga, D. S. Pengantar teori sekor pada pengukuran pendidikan. Besbats, Jakarta.. (1992).

[7] Camilli, G., & Lorrie, A. S. MMSS Methods for identifying biased test items. Thousand Oaks, CA, Sage Publication. (1994).

[8] Hambleton, R.K. & Swaminathan, H. (1985). Item response theory. Boston, MA: Kluwer Inc.

[9] Hullin, C. L., et al. (1983). Item response theory: Application to psychological measurement. Homewood, IL: Dow Jones-Irwin.

[10] Jahya Umar. Item Banking. In Advances in measurement in educational research and assessment. Geofferey N. Masters & John P. Keeves. Amsterdam, Pergamon (pp. 207-219). (1999).

[11] Mahmud, J . Item response theory: A basic concept. Educational Research and Reviews. Vol. 12(5), (pp. 258-266) 10 March, 2017, http://www.academicjournals.org/journal/ERR/article-full-text-pdf/53ED36963043, doi: 10.5897/ERR2017. 3147. (2017).

[12] Mislevy, R. J. & Bock, R.D. Bilog 3, Item analysis and test scoring with binary logistic models, (Second Edition). Mooresville: Scientific Software. Inc. (1990).

[13] Guyer, R., & Thompson, N. Item response theory parameter recovery using X calibre™ 4.1. Technical Report, August, 2011. Assessment Systems Corporation. https://www.assess.com/docs/Xcalibre_4.1_tech_report.pdf. (2011).

[14] Sahin, A, & Anil, D. The Effects of test length and sample size on item parameters in item response theory. Educational Sciences: Theory & Practice. 17(1). (pp. 321-335). http://www.estp.com.tr/wp-content/uploads/2016/12/ESTP-2017-0270.pdf, doi: 10.12738/escp.2017.1.0270. (2017).

[15] Akour, M., & Al-Omari, H. Empirical investigation of the stability of IRT item-parameters estimation. International Online Journal of Educational Sciences. 5(2). (pp. 291-301). https://eis.hu.edu.jo/deanshipfiles/pub106314725.pdf (2013)

[16] Hambleton, R., K. Prinsiples and selected applications of item response theory. In R. L. Linn (Ed). Educational Measurement (3rd ed, pp. 147-200). New York, NY: Macmillan. (1989).

[17] Tsutakawa, R. K & Johnson, J. C. The Effect of uncertainty of item parameter estimation on ability estimates. Psychometrika-Vol. 55, No. 2, (pp. 371-390). https://www.researchgate.net/profile/Jane_Johnson6/publication/24062923_The_effect_of_uncertainty_of_item_parameter_estimation_on_ability_estimates/links/0deec51e6c298222cd000000.pdf. (1990).

[18] Foley, B. P. Improving IRT parameter estimates with small sample sizes: Evaluating the efficacy of a new data augmentation technique. A Dissertation, University of Nebraska, http://digitalcommons.unl.edu/cgi/viewcontent.cgi?article= 1075&context=cehsdiss. (2010).

[19] Kang, T & Nancy S. Petersen, NS. Linking item parameters to a base scale. Asia Pacific Education Review. June 2012, Vol. 13, Issue 2, pp 311–321. http://link.springer.com/article. doi:10.1007/s12564-011-9197-2. (2012).

[20] Ping Chen et al. Online calibration methods for the dina model with independent attributes in CD-CAT. Psychometrika, April 2012, Vol. 77, Issue 2, pp 201–222. http:// http://link.springer.com/article/ doi: 10.1007/s11336-012-9255-7. (2012).