Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.10, No.30, 2019

Test Length and Sample Size for Item-Difficulty Parameter Estimation in Item Response Theory

Amen Valentine Uyigue Ph.D* Matilda Uvie Orheruata Ph.D
Department of Educational Evaluation and Counselling Psychology , Faculty of Education, University of Benin, Benin City, Edo State, Nigeria

Abstract
The study investigated test lengths and sample sizes in the accurate and stable estimation of item-difficulty parameter in the Item Response Theory (IRT) One Parameter Logistic Model (1PLM). Real data of students that sat for the June/July 2015 Economics Multiple-Choice Examinations in Edo State was obtained from the National Examinations Council (NECO), Nigeria. The statistical population of examinees were 5,158 and the test length 60. Sample sizes of 200, 500, 1000, 2000 and 5000 were randomly drawn from the population with replacement; these samples were each paired with test lengths of 10, 20, 30 and 50.all amounting to 20 statistical conditions (5 sample sizes× 4 test lengths)The parameter estimates were generated using the eirt - Item Response Theory Assistant for Excel. The generated item-difficulty parameter using the entire population was assumed to be the true parameter value against which others were compared, using the Root Mean Square Error (RMSE) as an evaluative criteria. The acceptable RMSE was ≤ 0.33. Conclusion reached was that for an accurate item-difficulty parameter estimate in the 1PLM at least a test length of 10 and sample size of 1000 is required.
Key words: Test-Length, Sample-Size, IRT, Difficulty-Parameter, Logistic Model.
DOI: 10.7176/JEP/10-30-08
Publication date:October 31st 2019

1. Introduction

Item Response Theory (IRT) also known as Modern Test Theory (MTT) is in the class of the Latent Trait Theory (LTT), it is a psychometric framework for item analysis and test development. It is a theory that puts item quality as well as the examinees abilities into considerations, when evaluating the psychometric properties (difficulty, discrimination and guessing parameters) of items in a scale.

IRT item parameter estimation entails a complex mathematical computation; though the use of computer soft-wares has reduced the rigour associated with the computation. However different available soft-wares use different estimation techniques, but the issues as noted by psychometric researchers is what constitute adequate number of items (test length) and sample size for an accurate parameter estimation. Hambleton, (1989) asserted that, test length and sample size needed for an accurate IRT item-parameter estimation is difficult to determine.

In IRT measurement framework there are at present three popular model of the dichotomous response category: the One- Parameter Logistic Model (1PLM) , Two-Parameter Logistic Model (2PLM) and Three-Parameter Logistic Model (3PLM) depending on the number of parameters (discrimination, difficulty and guessing) that is of interest

In some researches or under some investigative situations the interest of the researchers may be just on determining the difficulty level of items without regard for the discrimination and guessing as in the "Rasch Tradition". In a study conducted by Stone (2003) he reported that sample size is a major factor in obtaining stable parameters estimates when the Rasch model/1PLM is to be fitted to a data set, however, sometimes large numbers of examinees may not be available most especially in small scale testing as in the administration of the teacher-made-test. This condition should not prevent psychometricians from benefiting from the gains of IRT. Therefore what should be the minimum test length and sample size from an empirical point of view using real test data for accurate and stable difficulty-parameter estimation?

In this study the researchers deem it fit to empirically sample different test lengths against varying sample sizes in the estimation of the difficulty- parameters under the 1PLM.

Many psychometric researchers have published works on the effect of sample size and test length on the psychometric properties (Discrimination, Difficulty and Guessing Parameters) of items as well as the ability parameter of examinees, for example:

Stone (2003) did a study to determine the effect of sample size on the accuracy of item-difficulty parameter. In that study he had sample sizes ranging from 10 to 3,000 taken from an examinee population of 3,173. The samples were randomly selected into estimation conditions, the WINSTEPS statistical software was used in estimating the item-difficulty parameters. The estimated item-difficulty parameters attained an acceptable value and begin to converge only when the sample size got up to 500.

Akour, and AL-Omari (2013) conducted a similar study in Jordan they used sample size of 200, 500, 1000, 5000, 10000, and 20000 against test lengths of 15, 30 and 60. Data used in the study was an operational Mathematics data from a test that was conducted by the Ministry of Education in Jordan; about 40,000 testees

took the test. The 3LP model was fitted and they concluded that a test length of greater than 15 and sample size greater than 500 are needed for an accurate and stable item-difficulty parameter estimate.

Custer (2015) in a study examined item-difficulty parameter estimate with 40 items across various samples sizes. In the study 3,000 examinees were simulated with ability level that was normally distributed. Item-difficulty parameter that was estimated using the entire 3,000 examinees serves as the true item parameters estimates, while samples sizes of 100, 200, 300,... 1000 were randomly selected to make up 10 replications. The WINGEN statistical IRT software was used in estimating the item-difficulty parameter and he reported a sample size of 500 as the minimum requirement for stable parameter estimate.

Sahin, and Anıl (2017) conducted a study in which they considered various sample sizes (150,250,350,500,750 1000, 2000, 3000 and 5000) against different test lengths (10, 20 and 30) they concluded that both sample size and test length are important factors to consider in IRT item-parameter estimation and that sample sizes of 250, 350, 500 and 750 examinees can be used but it depends on test length, they presented a trade-off between sample size and test length, they concluded that a sample size of 150 is just okay for the estimation of the difficulty- parameter in the 1PLM irrespective of test lengths (10,20 or 30).

In a study conducted by He and Wheadon (2017) in which they investigated the effect of sample size on parameter estimates using the partial credit model revealed a trade –off between sample size and the accuracy of parameter estimation and they equally came to the conclusion that accuracy of item parameter estimation is a function of sample size.

2. Research Question:

What should be the acceptable test length and sample size for the estimation of Item-Difficulty Parameter in the IRT 1PLM?

3. Methodology:

The study adopted the Survey Research Design, the population comprised of the twenty three thousand two hundred and fifteen (23,215) examinees who sat for the National Examinations Council (NECO) Senior School Certificate Examination (SSCE) in Economics objective test paper III that was conducted in June/July 2015 in Edo State, Nigeria. The statistical sample was five thousand one hundred and fifty eight (5,158), the sample are the examinees that resounded to "Type C" option among the four available types A, B,C, and D.

Responses by examinees within the sample were randomly assigned with replacement to groups of 200, 500, 1000, 2000 and 5000 respectively; hence there were five sample sizes. The instrument contained sixty (60) items, for the purpose of the investigation the items were randomly selected with placements into groups of 10, 20, 30, and 50. Each of the sample size is paired with each test length; this amounted to twenty (20) statistical conditions, five sample sizes and four item lengths (5×4),as shown in table-1below.

Table 1: Test Lengths and Sample Sizes Distributions for Analyses

Model	Sample Sizes						
1PLM	Test Lengths	200	500	1000	2000	5000	**Total**
	10	1	1	1	1	1	5
	20	1	1	1	1	1	5
	30	1	1	1	1	1	5
	50	1	1	1	1	1	5
	Total	4	4	4	4	4	**20**

Estimates from the statistical sample and test length were treated/ assumed to be the true parameter values. In the report of Sawminathan, Hambleton, Sireci, Xing, & Rizavi, (2003) estimating the item-parameters using the entire population of examinees will produced the true item parameters .These values were compared with what was obtained in other combinations and analyses, all parameter estimations were done using the *eirt - Item Response Theory Assistant for Excel* statistical software by Germain, Valois, & Abdous, (2007).

The differences between the true parameter values and values obtained from other sub-samples were seen as the effect of sample sizes and test lengths. The Root Mean Square Errors (RMSEs) statistics was adopted in making this comparison. The computational formula for RMSE is presented;

$$\text{RMSE} = \sqrt{\frac{\sum_{k-1}^{k}(\delta i - \mathrm{T}i)^2}{K}}$$

Where: δ_i is the estimated item parameter while Ti represents "true" item parameter, and k is the test length.

The smaller the RMSE the closer the estimated parameter values are to the true parameter values and better estimates. In order to determine the feasible sample size and test length, Rudner, (1998) opined that RMSE $\leq$ 0.33, which corresponds to the classical reliability value of 0.90, is taken as the criteria for minimum feasible sample size for that particular test length and IRT model. Though, Han, Kolen, & Pohlmann (1997) asserted that

a RMSE less than 0.6 were considered small and equally considered as okay. The RMSE ≤ 0.33 was adopted in evaluating the results of this study, since it met the criteria for both, though a condition that appears to be more stringent.

4. Results

Table 2: RSME Distributions Across Test Lengths and Sample Sizes

Model	Sample Sizes					
1PLM	Test Lengths	200	500	1000	2000	5000
	10	0.767	0.495	0.211	0.171	0.091
	20	0.654	0.525	0.193	0.148	0.067
	30	0.722	0.483	0.151	0.137	0.063
	50	0.704	0.432	0.176	0.151	0.030

Table 2 contained the analysed results obtained under the various statistical conditions of test lengths and sample sizes, from the results; a sample size of 200 did not yield an acceptable RMSE when combined with test lengths of 10, 20, 30 and 50 ($RMSE_s > 0.33$). In the same vain a sample size of 500 did not yield an acceptable RMSE when combined with test lengths of 10, 20, 30 and 50 ($RMSE_s > 0.33$). However a sample size of 1000, yielded $RMSE_s$ less than 0.33, across various (10, 20, 30 and 50) test lengths, 10 = 0.211; 20 = 0.193, 30 = 0.151 and 50 = 0.176.Sample size of 2000, yielded $RMSE_s$ less than 0.33, across various (10, 20, 30 and 50), 10 = 0.171; 20 = 0.148, 30 = 0.137 and 50 = 0.151. While Sample size of 5000, yielded $RMSE_s$ less than 0.33, across various (10, 20, 30 and 50) test lengths, 10 = 0.091; 20 = 0.067, 30 = 0.063 and 50 = 0.030.

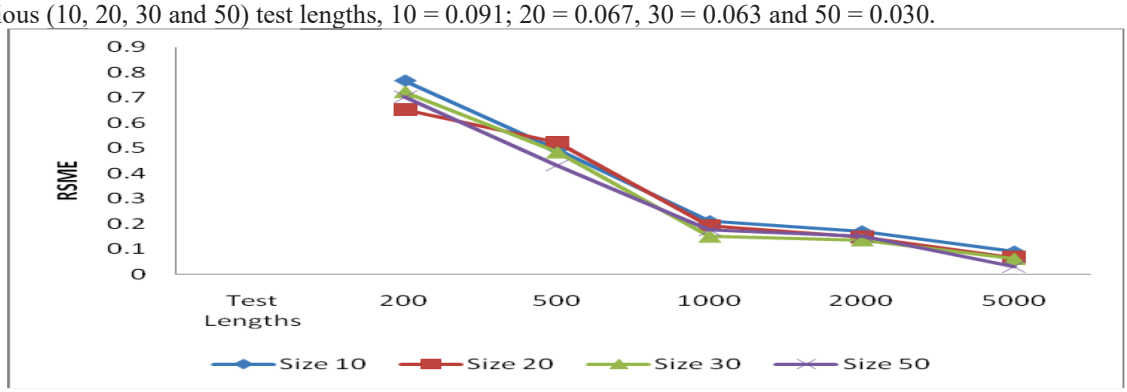

Figure 1: RSME of Item-Difficulty Parameter under Various Test Lengths and Sample Sizes.

Figure 1 shows the graphical representation of the result contained in Table 1, it can be observed that the lines representing the test lengths all went above the 0.33 RMSE in the y-axis under sample sizes 200 and 500 but less than 0.33 under 1000, 2000 and 5000 sample sizes.

5. Discussion of Findings

From the results presented above, it showed that for an acceptable, stable and accurate estimation of the item-difficulty parameter under the 1PLM a sample size of 1000 and test length as low as 10 can suffice. The finding contradicted the findings of Sahin, and Anıl (2017) who concluded that a sample size of 150 is good enough in combination with test of 10 and above. However there appears to be a point of agreement with respect to test length and again the trade-off between test length and sample size.

The findings from the study agree with the finding of stone (2003) who reported a convergence or stable parameter estimate only after the sample size attained a high of 500. Also the findings corroborate that of Custer (2015) who came to the conclusion that a sample size of 500 in combination with a test length 40 is needed for accurate item-difficulty parameter estimation. Though Custer attained a sample of 500 before arriving at this conclusion, the conclusion of 500 samples being good enough is not far from the fact that there exists a trade-off between test length and sample size, when a lesser test length is needed the sample size will definitely increase in order to obtained an accurate estimate.

The implication of the findings therefore is that for the IRT measurement framework to be used in test development there should be up to 1000 examinees and up to 10 items the possibility of 1000 examinees is a far cry from many operational situation, where the examinees may not be as many as 1000 therefore the IRT procedure is still limited to a large extent in its scope of applicability in test development.

6. Conclusion

From evidences gathered in the study the researchers therefore concluded that, when the focus in estimation is on the item- difficulty parameter alone as in the 1PLM, and a high level accuracy is desired, the sample size should be at least 1000 and test length at least 10. However a sample size of 500 with test length 10 can still yield an

acceptable item-difficulty parameter estimate, since the RMSE at these points still fall less than 0.6, a criterion provided by Han, Kolen, & Pohlmann (1997).

7. Recommendations

Arising from the findings of the study, the researchers recommended that when the IPLM is the model of choice employing the IRT framework a test length of 10 and sample size 1000 should be used in order to have high accuracy of the estimated difficulty parameter.

References

Akour, M & AL-Omari H (2013) "Empirical Investigation of the Stability of IRT Item Parameters Estimation" International Online Journal of Educational Sciences, 5 (2), 291-301

Custer, M. (2015). "Sample Size and Item Parameter Estimation Precision When Utilizing the One-Parameter "Rasch" Model": Paper Presented at the Annual Meeting of the Mid-Western Educational Research Association Evanston, Illinois October 21-24.

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Germain, S., Valois, P., & Abdous, B. (2007) eirt - Item Response Theory Assistant for Excel *(Freeware)*. Available online at: http://libirt.sf.net

Hambleton, R. K. (1989). Principles and Selected Applications of Item Response Theory. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 147-200). Washington, DC: American Council on Education and Macmillan.

Han, T., Kolen, M., & Pohlmann, J. (1997). "A Comparison among IRT True- and Observed-Score Equatings and Traditional Equipercentile Equating". *Applied Measurement in Education, 10*(2), 105-121.

He, C. & Wheadon, C.(2017) . "The Effect of Sample Size on Item Parameter Estimation for the Partial Credit Model" Center for Education Research and Policy. Retrieved August 2019 from https://www.semanticscholar.org

Lord, F. M. (1968). "An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-Parameter Logistic Model". *Educational and Psychological Measurement, 28*, 989-1020.

Mutasem, A., & Hassan, A-O. (2013). "Empirical Investigation of the Stability of IRT Item-Parameters Estimation": International Online Journal of Educational Sciences, 5 (2), 291-301

Rudner, L. M. (1998). "*An On-Line, Interactive, Computer Adaptive Testing Tutorial*." Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm

Sahin, A., & Anıl, D. (2017). "The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory". *Educational Sciences: Theory & Practice*, *17*, 321–335. http://dx.doi.org/10.12738/estp.2017.1.0270

Stone, M. (2003). "The Effect of Sample Size on Rasch/IRT Parameters Using Dichotomous Items". Retrieved April, 2019 from https://www.ncbi.nlm.nih.gov/pubmed/14757991

Swaminathan, H., Hambleton, R., Sireci, S., Xing, D., & Rizavi, S. (2003). *Small Sample Estimation in Dichotomous Item Response Models: Effect of Priors Based on Judgmental Information on the Accuracy of Item Parameter Estimates*. Newtown, PA: Law School Admission Council.

Dr. Amen Valentine Uyigue was born in Benin City Edo State, Nigeria .He is a lecturer in the Department of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin. He obtained the degrees of Bachelors of Science Education Bs.c (Ed.) Education Economics and Statistics in 2008, Maters (M.Ed) educational measurement and evaluation in 2012 and Doctor of Philosophy (PhD) educational measurement and evaluation in 2017 from the University of Benin, Benin City, Edo State, Nigeria.

Dr (Mrs.) Matilda Uvie Orheruata is a lecturer in the Department of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin. She holds a PhD Degree in educational measurement and evaluation from the University of Benin, Benin City, Edo State, Nigeria. Her research area is in Psychometrics and Educational Programme Evaluation.