Determining Test Length Precision for Economics Testing: The Implementation of IRT Model for Classroom Assessment

Friyatmi[1], Djemari Mardapi[2], Haryanto[3]

[1] Universitas Negeri Padang, Padang, Indonesia, 🖂 friyatmi@fe.unp.ac.id

[2] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia, 🖂 djemari@uny.ac.id

[3] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia, 🖂 haryanto.@uny.ac.id

Abstract

The Item Response Theory (IRT) has been used extensively in the test analysis to produce the estimation of item parameters and the ability of tester. However, test length and sample size can affect parameter estimates based on the IRT approach. The purpose of this study was to analyze the effect of sample size and test length on the stability of parameters estimation. This study was conducted using two stages of simulation and real testing in economics for senior high school students. The data were analyzed using the WINGEN for data generation and the BILOG for item parameter estimation and ability parameter estimation. The sample size were varied from small to large samples, while the length test were short tests (20 items) and long tests (> 20 items). The results showed that 1) sample size had an effect on the stability of the item parameters, where the increasing of the sample size impact on the raise of item parameter estimation, 2) the test length effect on the stability of the ability parameters, the rise of number of items effect on the increases of ability parameter estimation. Rasch models (1-PL IRT) can be applied by the teacher in conducting classroom assessment with test length 20 and has sufficient reliability.

Keywords: IRT, test length, sample size.

Introduction

The evaluation of learning outcomes plays an important role in enhancing the quality of education. A series of tests are usually chosen by the teacher in order to measure the student achievement in the classroom assessment. There are various type of tests, but multiple choice tests more often chosen because it offers simplicity in correction process. Although multiple choice items belong to the traditional test category, but behind their weaknesses this test is suitable for developing computer-based tests with high reliability (Conley, 2015). In the digital era, developing of computerized testing, various digital tools for assessment (Battro & Fischer, 2012) and automatic scoring test software (Mislevy, Behrens, Dicerbo, & Levy, 2012) will grow further by offering various efficiency to the testing process. This digital tests are not only used for large-scale tests but also have been developed for classroom assessment. Some studies show that the use of digital test is effective for the classroom assessment (He & Tymms, 2005 ; Hwang & Chang, 2011 ; Nicol, 2007 ; Scibinetti, Tocci, & Pesce, 2011). One important point in developing a test for classroom assessment is what is the proper number of items (test length) in a test? Test developers should know the minimum test length needed for an qualify test.

The determination of the adequate test length relating to the model used for test analyzing. There are two models that can be used in analyzing item test, that is classical test theory and item response theory (IRT). The use of classical test theory has been very familiar to the teachers because of its simple assumptions, easier measurement processes and supported by many simple user friendly computer applications. However, the use of classical test theory have a number of shortcomings because it is group group dependent and items dependent (Hambleton, Swaminathan, & Rogers, 1991). The application of IRT have been used widely in the practice of education measurement to cover the disadvantage of classical test theory. IRT has several advantages: 1) the item characteristics do not depend on the examines; 2) the score described by the examines does not depend on the whole test characteristics, 3) the model that emphasizes more on item level than the test, 4) does not require strict parallel tests to estimate reliability, and 5) describes a decision size for each ability score is that there is a functional relationship between examinee and their level of ability.

The use of IRT in test analysis not only produce item parameter estimation but also provide information on the estimated ability of the examines. IRT provides a more stable statistical process for estimating the characteristics of both items and the examines (Brennan, 2006). It also has the ability to define how these characteristics interact in describing item abilities and tests characteristic. Based on this theory, IRT contains two parameters, namely the item parameters and examines parameters. Ability parameter (ϴ) located in the interval of $-\infty \leq \theta \leq \infty$ (Kolen & Brennan, 2004) and scaled close to normal distribution with mean 0 and standard deviation 1. But in practice a person's ability lies between -3.0 to 3.0, even Naga (1992) states that a person's ability is practically between -4.0 to +4.0. Item parameters are indicated through a suitable logistic model. There are several measurement models that are commonly used in the IRT for items analysis. The selection of the right model will reveal the true characteristics of the test data as a result of the measurement. In the IRT for dichotomous test data, three logistic models can be used, namely one parameter logistic (1-PL), two parameters logistic (2-PL), and three parameters logistic (3-PL). The probability of the examinees to correctly answer an item in the one parameter logistic model is determined by one item characteristic, namely the item difficulty index (b). Two-parameter logistic model, the probability of the examinees to correctly answer a question item is determined by two item characteristics, that is the difficulty index (b) and a discrimination index (a). While the three-parameter logistic model is determined by three item characteristics, namely the item difficulty index (b), the discrimination index (a), and the pseudo guessing parameter (c).

There are several factors that may affect parameter estimation based on the IRT model, including estimation of the IRT model, test length, and sample size (De La Torre, Hong, & Deng, 2010). Test length related to the number of items and sample size refer to the number of examinee in a test. Chuah, Drasgow, and Luecht (2006) suggest that sample sizes as small as 300 respondents are adequate for estimating ability for adaptive testing. Furthermore, (Fitzpatrick and Yen (2001)) requires more than 200 sample size to obtain accuration item parameter estimates when using 2-PL IRT. Research show that the use of different IRT models resulted in the different of stability parameter estimation (Baur & Lukes, 2009). Barnes and Wise (1991) suggested the 1-PL IRT model as the most effective IRT model and use it in many measurement context. Moreover, Svetina et al. (2013) noted in the 1-PL estimation model, the accuracy of parameter estimation decreased when using small sample size and test length, indicated that there is an effect of sample size on the stability of parameter estimation. The question arise to what is the proper sample size that would produce the stability of the parameter estimation? Sireci (1992) suggested a minimum IRT analysis within a sample size of 200 people. The number of sample size increased as the parameter estimations utilized increase. This means that the use of 2-PL and 3-PL models requires a sample size of more than 200. In addition to the sample size, the selection of the model in parameter estimation is also related to the accuracy of the parameter estimation. Based on these studies, it suggests that the 1-PL IRT model is the appropriate model for classroom assessment because of the limited number of students in the class.

This study was designed to examine some effects of test length and sample size that produces the most accurate results of parameter estimates and determining the proper test length for classroom assessment. The purpose of this research was to describe; 1) the influence of sample size on the stability of item parameter estimation, 2) the effect of number of item or test length on the stability of examinee parameters (ability) estimation. A simulation study was conducted to evaluate the accuracy of test length and sample size using 1-PL IRT model. Then, a real testing applications must be address to generalize of this study.

Methods

This research was carried out through some simulation and real testing in economics for senior high school students. There were two simulations conducted to examine the effect of the test length and sample size factors using IRT model on item parameter and ability estimation. The sample sizes used were varied for small and large samples. The length of the test was also chosen for short tests (20 items) and long tests (> 20 items). Variations in the number of items for a long test are carried out at amounts close to 40 items, that is the number of items in the national exam for economic subject. All simulations carried out use normal distribution and replication is done five times. Table 1 illustrated the simulation design that was carried out.

Table 1 The Simulation Design

No.	Variety	Test Length (n)	Sample Size (N)	Model
1.	Simulation1	40	40	1-PL
			400	
			2000	
2.	Simulation 2	20	1000	1-PL
		40		
		60		

The simulation data were generated by WinGen Software, while Bilog program were utilized to analyze the item parameter and ability estimation. Parameter estimation consistency can be seen from the true parameter correlation value and RMSE estimated parameters and values. The higher the parameter estimation correlation, the higher or more stable item parameter estimation of examinees. The opposite is indicated by the value of RMSE. The lower the RMSE value, the higher the stability of item parameter estimation and examinees.

The consistency of estimation parameters can be seen from the true correlation of parameters and parameter estimation and the RSME (Root-Mean-Square Error). The higher parameter estimation correlation shows the stable or the accuracy of parameter estimation. The opposite is shown by the RMSE. The lower RSME evince the higher stability of parameter estimation. The correlation of the ability parameter estimation of the participants was analyzed using Microsoft Excel program based on the average of the 5 replications performed.

As a comparison, the real testing were held in the economics subject using varied sample sizes (40, 100, and 200 testee) and three test lengths (20, 40 and 50 items) to the senior high school students. The purpose of this test was to analyze the accuracy of parameter estimation of IRT model for the classroom assessment.

Results and Discussion

1st Simulation

The Effect of Sample Size on The Accuracy of Item Parameter Estimation

Effect of sample size (N) on the stability of item paramater estimation can be seen from the correlation coefficient between true item parameter and item parameter estimation (difficulty index (b)) and based on RMSE. Table 2 display the results of data analysis using the 1-PL IRT model.

Table 2 Correlation Coefficient for Item Parameter and RSME on 1st Simulation

	Correlation of True and Estimate Paramater			RSME		
	N = 40	N = 400	N = 2000	N = 40	N = 400	N = 2000
Rep1	0950	0994	0999	0.409	0.125	0.063
Rep2	0930	0995	0999	0.419	0.121	0.083
Rep3	0901	0994	0999	0.630	0.125	0.067
Rep4	0928	0995	0999	0.440	0.118	0.074
Rep5	0923	0995	0999	0.479	0.116	0.079
Average	0926	0995	0999	0.475	0.121	0.073

The simulation results show that when the sample size was multiplied, the correlation parameter estimation level difficulty was also getting higher. It shows that there is a positive relationship between sample size and item parameter estimates stability. The larger the sample size, the more stable the item parameter estimation is. The relationship can be described in the following graph.

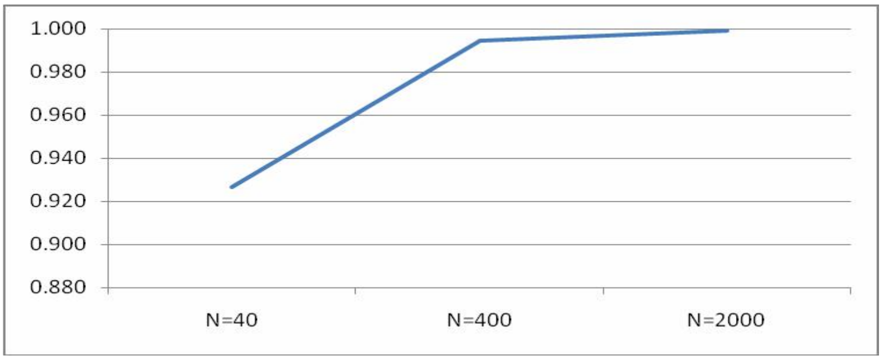

Figure 1 Effect of Sample Size to The Accuracy of Item Parameter (b)

The Figure 1 indicated that there is an increase in the number of examinees from 40 to 400, there is also an increase in the correlation of estimated item parameter (difficulty index) which is quite high. However, when the number of examinees added from 400 to 2000 people, the increase in the correlation of item parameter estimation tends to be slightly. This implies that the estimation of item parameter using 1-PL IRT model does not need the big sample size because the results were not significant.

An alternative method to observe the stability of parameter estimation is through the RMSE. Data on the Table 2 shows that the more sample size will be lower the error estimation parameter. The lower RMSE shows the more accurate item parameter estimation. This shows that there is a negative relationship between the sample size and the item parameter estimation error.

Based on the first simulation, it can be concluded that the increase in number of testee will result in increasing the accuracy of parameter estimation. On the other hand , the smaller samples size will result in the lower accuracy of parameter estimation.

The Effect of Test Length (n) on the Examinees Parameter (θ)

The effect of the number of items (n) on the accuarcy of the examinees parameter can be observed from the correlation coefficient between true parameters tetha and the estimation of tetha.

Table 3 The Coefficient Correlation of True and Estimation Tetha

Korelasi True dan Estimasi Tetha			
	n=20	n=40	n=60
Rep1	0.879185	0.822803	0.947381
Rep2	0.870069	0.852523	0.846423
Rep3	0.87505	0.908037	0.949143
Rep4	0.871546	0.877538	0.933597
Rep5	0.877038	0.924613	0.947856
Rerata	0.874578	0.877103	0.92488

Table 3 shows that the more the number of items impact on the higher correlation parameter estimation examinees (tetha). This shows that there is a positive relationship between the test length and the stability of the examinees parameter estimation. The increase of number of items from 20 to 40 results a slight increase in ability estimated. However, when the number of items was 60 items, the correlation of the ability estimation increased quite sharply. Based on the results, it can be concluded that the increase in the number of items will result in increasing the accuracy of the examinee estimation. On the other hand, the smaller test length will effect on the lower the stability of the ability parameters.

2nd Simulation

Influence of Sample Size and Test Length on Item Parameters

The raise of number of examinee on each test length impact on the increasing coefficient correlation of the item parameter. It means the accuracy of the item parameter estimation will be achieved on the higher sample size. In contrast, the rise of the test length in a specific sample size will not effect the item estimation. For instance, when the sample size is 500, the correlation coefficient of item parameters tends to be relatively the same in each test length. It happens on all levels of test length. It could be happened because the change in test length and sample size is infrequently

Table 4 Correlation Coefficient for Item Parameter

Correlation of True and Estimate Parameter Item			
	N50	N250	N500
n30	0.960	0.993	0.995
n40	0.957	0.990	0.995
n60	0.954	0.991	0.996

Based on simulation 2, It can be concluded that the increasing of the sample size result the accuracy of item parameter estimation using 1-PL IRT. On the other hand, the increase of test length does not significantly affect the stability of item parameter estimation.

Effect of Sample Size and Test Length on Examinees parameter

The result shows that the increase of number of items impact on the raise of the ability parameter correlation coefficient for all variations sample size. Diversely, the increase of sample size tends to be meaningless to the ability parameter correlation coefficient.

Table 5 The Coefficient Correlation of True and Estimation Tetha

Coefficient Correlation Ө			
	n30	n40	n60
N50	0.9197	0.94254	0.95078
N250	0.9160	0.93419	0.94939
N500	0.9101	0.93452	0.95114

Based on simulation, it can be concluded that the longer test result the higher accuracy of the parameter ability. However, the more the number of examines in a test, it turns out that it does not significantly affect on the stability of the ability parameter estimation.

Real Testing Analysis for Classroom Assessment in Economic Learning

In this study, the three economics test was administered to the senior high school student. It were tested with varied sample sizes (40, 100, and 150 testee) and three test lengths (20, 40 and 50 items).

Estimation of Items Parameter

As shown in Table 6, sample size more influence the differences between parameter estimation rather than test length. The data show that the estimation of level difficulty will increase when the samples size grow up. Differently, the rise of the test length does not encourage the increases of the level difficulty. The real testing prove that it is relevant to the simulation conclusion

Table 6 Item Parameter Estimation on Real Testing

	Parameter Estimation (b)		
n	N40	N100	N150
n20	1,968	1,897	1,982
n40	-0,096	0.145	0.188
n50	-0.158	0.099	0.162

Ability Parameter Estimation

Data in The Table 7 show that the estimation of ability tend to increase when the test length raise. The results of this analysis are relevant to the simulation results which conclude that the longer a test will result the higher ability estimation.

Table 7 Ability Estimation based on Real Testing

n	Ability Estimation		
	N40	N100	N150
n20	-0.0015	0.0468	0.0497
n40	0.0014	-0.0015	0.0161
n50	0.0950	0.0437	0.0558

Based on real testing, the item parameter estimation are more stable when involving test lengths around 40-50. But to determine how many items are right in class assessment must be supported by reliability information.

Table 8 Reliability

n	Reliability		
	N40	N100	N150
20	0.72	0.61	0.58
40	0.94	0.93	0.93
50	0.95	0.95	0.94

The reliability test tends to be low when using the 20 items in the test. The addition of the sample size resulted in decreasing reliability. For classroom assessment, where it usually consist of 30 to 40 students, the one parameter IRT model with test length 20 items can be an alternative analysis because it has sufficient reliability (reliability index minimal 0.7). However, the test length 40 items has been an ideal model in producing the accurate parameter estimation and a high reliability.

Based on the result, it can be seen that the number of examinee (sample size) affects the stability of the parameter estimation. The more sample size, the higher the stability of item parameter estimation. This is evidenced by the high correlation of parameter estimation, and low error estimation. However, when the sample size is small, for instance below 100 people, then the stability of item parameter estimation also tends to be low. The relationship between sample size and stability of parameter estimation is positive, as revealed by Custer (2015) that there is a positive relationship between sample size and item parameter estimation precision using IRT, which means that when the sample size is large the parameter estimation precision will be high, and vice versa .

The accuracy of sample size used in estimating item parameter is also very dependent on the IRT model (Delucchi, 2004). The use of a small sample size in the 1-PL model may not be too problematic, but the result is different if using the 3-PL model. The use of a small number of samples with the 3-PL model resulted in the wingen program being unable to complete the estimation, because it found several outputs that were empty when the estimation used a small sample size. This proves the use of a number of samples that are slightly unsuitable using the 3-PL estimation model. Tang, Way, and Carey (1993) suggested the use of a large sample size when estimating using a 3-PL model that is at least 1000 people so that the resulting parameter estimation are more accurate and stable. Conversely, small size sample usage is still able to provide stability in parameter estimation when using the 1-PL model. As expressed by Stone and Yumoto (2004) that the use of Rasch models (1-PL) allows for parameter estimation using a small number of samples, even with a sample of 30 people. In this simulation, with the least sample size of 40 people and the analysis model used is 1-PL, the item parameter estimation can still be generated with a correlation value of 0.926 but with a high RMSE value of 0.475. When using a sample size of 250, the RMSE value becomes much lower. This result is in accordance with previous studies which require a minimum sample size for a more stable IRT parameter estimation is 200 (Sireci, 1992). But at least in the context of classroom assessment, Rasch model can be used as an alternative for teachers to estimate parameters of test items that are IRT-based rather than just using classical test theory by making the same package test for several classes.

The use of different test lengths in this simulation also has a different effect on the stability of the participants' ability parameter estimation. Where the length of the test increases, the SE value will decrease and the correlation of item parameter estimation will be higher or more stable. With variations in the number of items 20, 30, 40, 60, and 80 items in this simulation, the stability of the parameter estimation of the highest item difficulty level generally occurs in the number of questions around 40 to 60 items. However, when the number of items was 80 items, the stability of the parameter estimation was lower than the number of items 60.

Although the sample size influences the item parameter estimation, in this study there is not enough evidence that the size of the sample has an effect on the ability of participants. Likewise, the addition of the test length does not significantly affect the stability of item parameter estimation.

Conclusions

Based on the results of the simulation and real testing, the following conclusions can be drawn. 1) Samples size affect on the stability of the item parameters, where the increase of the sample size impact on the accuracy of item parameter estimation, and vice versa. 2) The test length effect on the stability of the ability parameter. The more raise of number of items results the more accurate the ability estimation. On the contrary the lower the number of items impact on the more unstable the ability estimation. Rasch models (1-PL IRT) can be applied by the teacher in conducting classroom assessment with test length 20 and has sufficient reliability. However, the number of items 40 is the proper test length because it has high accuracy of the parameter estimation and higher reliability.

References

Barnes, L. L., & Wise, S. L. (1991). The utility of a modified one-parameter irt model with small samples. *Applied Measurement in Education*. Vol. 4, No. 2, 143-157.

Battro, A. M., & Fischer, K. W. (2012). Mind, brain, and education in the digital era. *Mind, Brain, and Education*. Vol. 6, No. 1, 49-50.

Baur, T., & Lukes, D. (2009). An evaluation of the irt models through monte carlo simulation. *UW-L Journal of Undergraduate Research*. Vol. 12, No., 1-7.

Brennan, R. (2006). *Educational measurement*. Westport: Greenwood Publishing Group.

Chuah, S. C., Drasgow, F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. *Applied Measurement in Education*. Vol. 19, No. 3, 241-255.

Conley, D. (2015). A new era for educational assessment. *education policy analysis archives*. Vol. 23, No., 8.

Custer, M. (2015). Sample size and item parameter estimation precision when utilizing the one-parameter" rasch" model. *Online Submission*. Vol. No.

De La Torre, J., Hong, Y., & Deng, W. (2010). Factors affecting the item parameter estimation and classification accuracy of the dina model. *Journal of Educational Measurement*. Vol. 47, No. 2, 227-249.

Delucchi, K. L. (2004). Sample size estimation in research with dependent measures and dichotomous outcomes. *American journal of public health*. Vol. 94, No. 3, 372-377.

Fitzpatrick, A. R., & Yen, W. M. (2001). The effects of test length and sample size on the reliability and equating of tests composed of constructed-response items. *Applied Measurement in Education*. Vol. 14, No. 1, 31-57.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park: Sage Publication Inc.

He, Q., & Tymms, P. (2005). A computer-assisted test design and diagnosis system for use by classroom teachers. *Journal of Computer Assisted Learning*. Vol. 21, No. 6, 419-429.

Hwang, G.-J., & Chang, H.-F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*. Vol. 56, No. 4, 1023-1031.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking*. New York: Springer.

Mislevy, R. J., Behrens, J. T., Dicerbo, K. E., & Levy, R. (2012). Design and discovery in educational assessment: Evidence-centered design, psychometrics, and educational data mining. *JEDM| Journal of Educational Data Mining*. Vol. 4, No. 1, 11-48.

Naga, D. S. (1992). *Pengantar teori sekor pada pengukuran pendidikan*. Jakarta: Gunadharma.

Nicol, D. (2007). E-assessment by design: Using multiple-choice tests to good effect. *Journal of Further and Higher Education*. Vol. 31, No. 1, 53-64.

Scibinetti, P., Tocci, N., & Pesce, C. (2011). Motor creativity and creative thinking in children: The diverging role of inhibition. *Creativity Research Journal*. Vol. 23, No. 3, 262-272.

Sireci, S. G. (1992). *The utility of irt in small-sample testing applications.* Paper presented at the The Annual Meeting of the American Psychological Association 100th.

Stone, M., & Yumoto, F. (2004). The effect of sample size for estimating rasch/irt parameters with dichotomous items. *Journal of applied measurement*. Vol. 5, No. 1, 48-61.

Svetina, D., Crawford, A. V., Levy, R., Green, S. B., Scott, L., Thompson, M., . . . Kunze, K. L. (2013). Designing small-scale tests: A simulation study of parameter recovery with the 1-pl. *Psychological Test and Assessment Modeling*. Vol. 55, No. 4, 335.

Tang, K. L., Way, W. D., & Carey, P. A. (1993). The effect of small calibration sample sizes on toefl irt-based equating. *ETS Research Report Series*. Vol. 1993, No. 2, i-38.