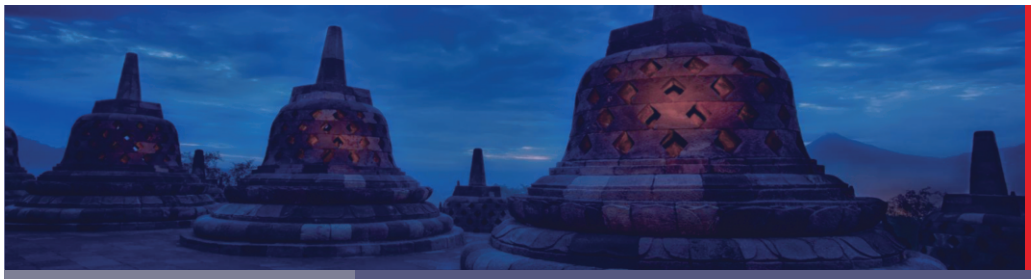

ISSN: 2407-1501

PROCEEDING

INTERNATIONAL CONFERENCE ON EDUCATIONAL RESEARCH AND EVALUATION (ICERE)

"Assessment for Improving Students' Performance"

May 29 – 31 2016
Rectorate Hall and Graduate School
Yogyakarta State University
Indonesia

Organized by:
Study Program of Educational Research and Evaluation
Graduate School, Yogyakarta State University
in Cooperation with Indonesian Educational Evaluation Association (HEPI),
and Center for Educational Assessment (PUSPENDIK) Ministry of Education and Culture

Proceeding
International Conference on Educational Research and Evaluation (ICERE) 2016

Publishing Institute
Yogyakarta State University

Director of Publication
Prof. Djemari Mardapi, Ph.D.

Address
Yogyakarta State University
ISSN: 2407-1501

Foreword of the Chairman

Assalamualaikum wr. wb.

Good morning ladies and gentlemen.

Praise be to Allah who has given abundant blessings so that we can hold this international conference.

This conference is aimed at improving the quality of assessment implemented in schools and other institutions. The quality of assessment determines students' ways of learning, so that it is hoped that the quality of education improves. Besides, this conference is a means of information exchanges in the forms of seminars dealing with results of research in educational assessment and evaluation. The expectation is that there is always improvement in educational assessment and evaluation methods, including in it is the instrument – both cognitive and noncognitive instruments.

The participants of this conference are the lecturers and teachers who teach educational assessment and evaluation, practitioners of assessment and evaluation, and researchers of assessment and evaluation. This conference can be held in cooperation with the Graduate School, Yogyakarta State University, Association of Educational Evaluation of Indonesia (HEPI), and Centre for Educational Research, Ministry of Education and Culture of Indonesia, supported by the Australian Council for Educational Research (ACER), Intel, Intan Pariwara Publisher, and many other institutions. For this reason, on behalf of the Organizing Committee, I would like to thank the Rector of Yogyakarta State University, Prof. Dr. Rochmat Wahab, M.Pd., M.A., and the Director of Graduate School, Yogyakarta State University, Prof. Dr. Zuhdan Kun Prasetyo, M.Ed., and all other institutions for their assistance and contribution that have made this conference possible. I would like to thank HEPI's Local Coordination Unit and all sponsors for supporting this conference and also all the audience for participating in this conference.

To the committee members, both in Jakarta and Yogyakarta, I would like to thank them for the hard work they have performed and for the togetherness so that this conference can be held.

Last but not least, we apologize for all the incoveniences you might encounter during this conference. Please enjoy the conference.

Wassalamu'alaikum wr. wb.

Prof. Djemari Mardapi, Ph.D.

Foreword of the Chairman of Himpunan Evaluasi Pendidikan Indonesia (HEPI)

Assalamu'alaikum Wr. Wb.

Indonesian Association for Educational Evaluation (HEPI) is a professional organization in education holding in the high esteem the principles of professionalism and knowledge development in the field of educational and psychological measurement, assessment, and evaluation. HEPI was established in November 19, 2000 in Yogyakarta, with a vision to become a professional organization that excels in the field of evaluation and measurement in education and psychology in Indonesia. Its mission is to develop up-to-date methodologies of evaluation, assessment, measurement, and data analysis in education and psychology, as well as studies of policies and technical implementation of the field for improving Indonesian education quality.

As a professional organization, HEPI brings together experts, practitioners and interested persons in the field of evaluation, assessment, and measurement of education, psychology and other social sciences. HEPI is open to anyone who has the interest the field with no restriction in terms of educational background and working experiences. Hopefully, through HEPI, members of the association can sustainably develop themselves as professionals. The existence of HEPI is also expected to contribute to the improvement of the quality of national education through research, consultancy, seminar, conference, publication, and training for members of the organization and for public audiences.

HEPI organizes annual workshop and conference in cooperation with the Regional Chapter of HEPI and universities. In 2016, for the first time HEPI organized **International Conference on Educational Research and Evaluation: Assessment for Improving Student's Performance** in May 29-30 2016 in Yogyakarta. This conference is jointly organized by HEPI and Yogyakarta State University and supported by the Center for Educational Assessment the Ministry of Education and Culture, Australian Council for Educational Research (ACER), INTEL Indonesia, and Intan Pariwara Publisher.

It is important to note that the choice of the HEPI 2016 conference theme is driven by the fact that the quality of our national education is still under expectation as shown by the results from School National Exam and international surveys conducted by some international agencies. HEPI believes that a number of factors contribute to the low quality of national education, including low teacher's knowledge and skills in classroom and school assessment. Therefore, improving the competence of teachers in classroom and school assessment is urgently required. In this context HEPI as a professional organization and individual members of the organization have to play an active role in improving teachers' competence in quality learning assessment.

In line with 2016 conference theme, HEPI invited two respected guest speakers , namely, Professor Geofferey Masters, Ph.D., Director of the Australian Council for Educational Research (ACER), who presented a paper on Assessment to Improve Student Competency and Professor Frederick Leung, Ph.D., from the University of Hong Kong, who delivered a paper on the International Assessment for Improving Classroom Assessment.

As a tradition, in 2016 conference HEPI organized two pre-conference workshops. The first workshop is on the conceptual introduction of Rasch model by Jahja Umar, Ph.D., senior lecturer at the Faculty of Psychology, State Islamic University Jakarta and the second workshop was delivered by Heru Widiatmo, Ph.D., researcher at American College Testing (ACT) Iowa, United States on Measuring Higher Order Thinking Skills (HOTS).

On behalf of HEPI, I would like to express my heartfelt gratitude to Rector of the Yogyakarta State University, invited speakers, resource persons, HEPI regional chapters, sponsors, speakers, participants, invited guests, and organizing committee who have worked hard in making this international coneference a success. Thank you very much for your participation and support and we are looking forward to seeing you in the next conference.

Last but not least, we hope that all of us get much benefit from this conference for enhancing Indonesian quality education through quality assessment.

Wassalamualaikum wr. wb.

Chairman,
BAHRUL HAYAT, Ph.D.

Table of Contents

THE ACCURACY OF TESTEES' ABILITY ESTIMATION OF THE ESSAY TEST AND TESTLETS IN MATHEMATICS THROUGH THE GRADED RESPONSE MODEL (GRM) APPLICATION

Purwo Susongko, Wikan Budi Utami
Science Education Departement, University of Pancasakti Tegal
Mathematics Education Departement, University of Pancasakti Tegal
(Kusumatirto@gmail.com)

Abstract - The purpose of this study are to find: (1) The accuracy of estimation of the testees' ability of mathematics achievement of the essay test and testlets through the Graded Response Model (GRM) Application, (2) The precision of estimation of the testees' ability of mathematics achievement of the essay test and testlets through the Graded Response Model (GRM) Application. This study was conducted in two stages, empirical and simulation. Empirical studies conducted as the mathematics achievement test of 277 Year X students SMA III Slawi on 2015. The testees' ability parameter of the estimation results of empirical study are used to generate true parameter of testees' ability. The data were generated from the parameters of the estimation results of the empirical study. The data were generated on the basis of the sample sizes of 500 testees and the numbers of items were 10. Replication is done 25 times, using WinGen 3.0 program. Item Parameters generated on the standard normal distribution. The parameters of testees' ability estimated by using a computer program Multilog Version 7:03. Precision in parameter estimation used to mean squared error (MSE). To assess the accuracy the coefficient of Pearson Correlation was calculated. The findings of the study are: (1) The accuracy of the estimation of testees' ability of mathematics achievement essay test is higher than testlets, (2) The precision of the estimation of testees' ability of mathematics achievement essay test is higher than testlets.

Keywords*: accuracy, precision, testees'ability, essay test , testlets*

I. INTRODUCTION

Test can use the collection of multiple items in a group are interconnected. A set of interrelated items which are part of a test commonly called testlets (Wainer, 1995). Testlest been used in the measurement in the field of language (Lee et al, 2001), TOEFL (Wainer & Wang, 2000), law school admission test (Wainer , 1991), medical school admission test (Zenisky et al, 2002), the measurement on a large scale (Ferrara et al, 1997) and attitudes toward racial identity (Fischer & Tokar, 1998). In the TOEFL test, testlets mainly used in the reading. Testlets also widely used in the world of public health and medical (Hamilton, Clayon B, et al 2015).

Reported by Zenisky, Hambleton, & Sireci (2002) that the item in one testlets not assume the character of local independence so that when these items are considered as items that are independent then it does not meet the requirements analysis to item response theory (IRT). Lee, et al. (2001) did equalize using polytomous item response theory, on a test composed of testlets. Lee, et al. (2001) showed that better psychometric equivalency using collection of items as a single response that is polytomous than items is considered as the independent items.

Many studies showed that the response of testlets to testlets more good in psychometric view when modeled in polytomous response that are multikategorik. Some experts have devised

theoretical support to the application of polytomous IRT on testlets scoring (Verhelst & Verstralen, 1997). Wainer, Bradlow & Wang (2007) developed a theoretical framework for the analysis of items test based testlest then called Testlets Response Theory (TRT). The theoretical framework shows that testlets can be analyzed using polytomous IRT. It provide the possibility the test items in an essay test with that in testlets that have many advantages .

There are many benefits when tests are administerated in testlets. In addition to the advantages of objective tests in general, testlet also has a scoring system that is multikategorik. Their similarity of IRT analyzes in essay test and testlets raises the question of the extent to which differences in the effectiveness of the testlet and essay test. Some studies do a comparison of both of the test, but is limited to non psychometric aspects of the classical theory approach so as not to provide information about the measurement accuracy of both of the test. By comparing the effectiveness in psychometric view will be known in accuracy when used as a instrument.

Purwo Susongko (2009) conducted a comparison of the effectiveness of the essay test and testlets and through graded response model application. The results showed that the empirical and simulation, the average value of the function information item on the essay test higher than the average value of the function item information in the testlets. This implies that the essay test is more effective than testlets from the informationzitem function is generated.

The study does not answer about the accuracy of the both test to estimate on the testees' ability. The accuracy of testees' ability estimation is parameter of the effectiveness in psychometric view because basically the test is to determine the testees' ability. Further, it needs research about differences in the accuracy of estimates of the testees' ability in an essay test and that in testlests. Based on the background, the problem in this research are as follows:

1. Is there a difference in the accuracy of estimation the testees' ability on mathematics achievement test in an essay test with that in testlets through the Graded Response Model (GRM) Application?
2. Is there a difference in the precision of estimation the testees' ability on mathematics achievement test in an essay test with that in testlets through the Graded Response Model (GRM) Application?

II. RESEARCH METHOD

The development of the instrument initiated by compiling mathematics achievement tests as mathemathics achievement test for Year X students of Senior High School. The instrument consists of 5 essay items. With the same test items,the essay tes was converted into 5 testlets. The number of items was adjusted to the material coverage and the time allocation which was 60 minutes. Mathematics achievement test was made by researchers in collaboration with the Mathematics teachers in SMA III Slawi. Both of the tests is admisnisterated on 277 students of class X in SMA III Slawi.

To obtain evidence of the content validity or content representation, the test papers were assessed through the expert judgment. To prove the construct validity and the unidimensional assumption of the tests developed, the data obtained from the tryout results were analyzed by using the structural equation modeling (SEM). The SEM employed to prove the construct validity and the unidimensional assumption was the confirmatory factor analysis (CFA) model. The program employed to analyze the data using this model was PRELIS to obtain the product moment correlation coefficients among items. The scoring of the two instruments employed the analytic method involving four criteria, namely scores 0, 1, 2, and 3. The answer to the previous stages affect the next stage, so that students who answered correctly on the first stage was given a score of 1 students who can answer correctly the whole stage, was given a score of 3 students who answer correctly the second stage, but the first one stage or in two stages of students answered incorrectly given a score of 0. scoring guidelines (rubric answer) created by step completion of each item and have been discussed with team (researchers and teachers). First rater and second rater respectively mathematics teacher who taught in SMA III Slawi Year X.

Testees'ability are estimated using a computer program Multilog Version 7:03. The program is selected because it is easy to use, easy to understand the reading results of the analysis. To determine the accuracy of the testees' ability parameter estimation of essay test and testlet conducted simulation studies. Empirical studies used to determine the value of the testees' ability parameter. Through simulation studies, the testees' ability parameter of empirical study are used

to generate the true score of testees' ability for each test . The data were generated on the basis of the sample sizes 500 testees and the numbers of items were 10. Replication is done 25 times, using WinGen 3.0 program. Item parameters were generated from the item through simulation using the standard normal distribution.

Simulation data used to estimate of the testees' ability. To determine the precision of the estimates used the mean squared difference between the true value and the estimates across replication , referred to as the Mean Squared Error (MSE). As for knowing the accuracy of the estimation used Pearson's r correlation between the true score and the estimates across replication.

With smaller of MSE average value of testees' ability estimation showing that the test form is precise, while larger values of MSE average value of testees'ability estimation showing that the test form is not precise. From correlation coefficient between true score and the testees' ability simulation results can indicate the extent to which the accuracy. Greater the mean of correlation coefficient between true score with the testees' ability simulation results show that more accurate estimates of testees' ability

III. RESEARCH FINDING AND DISCUSSIONS

The empirical data were collected through students' responses to the essay test and the testlets administered to 277 Year I students of the mathematics and science program (MIA) and Social studies Program (IPS) in SMA Negeri 3 Slawi. The test was carried out on May 25, 2015 at 8:00 to 10:00 pm. Construct validity test has been conducted on both the test by using Structural Equation Model (SEM) analysis. SEM analysis is used for the confirmatory validity analysis. Path diagram for the construct validity analysis results of the testlets can be seen in Figure 1, while of the essay test can be seen in Figure 2.

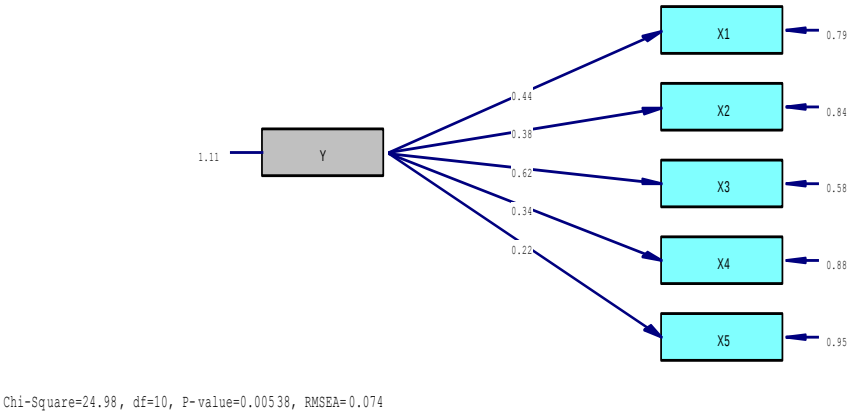

Figure 1. Path Diagram of construct validity analysis on testlets

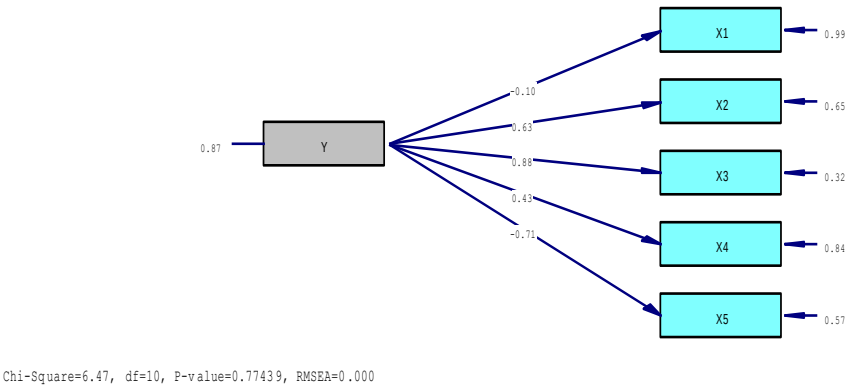

Figure 2. Path Diagram of construct validity analysis on essay test

Of the two pictures can be seen that the model used can be considered valid at least with the following considerations: (1) the value of the ratio chi-square of the degree of freedom of each is

2.498 and 0.647, the value of both are much lower than the cutoff values suggested by Wheaton (1977) that is equal to 5., (2) the value of RMSEA of essay test is 0,000 that is far from the maximum values suggested by Joreskorg & Sorbom (1996) that is equal to 0.05, while in the testlets is 0, 074, that the value not too extreme and is still considered to be rational when compared with those criteria.

The following is the item parameters of the estimation results of the empirical study of the testlet (Table 1) and the essay test (Table 2).

Table 1. Item Parameters of the Testlet by Empiric

Item Number	b 1	b 2	b 3	a
1	-1.72	1.07	1.32	0.73
2	-1.10	0.00	1.10	1.10
3	-9.46	-6.19	1.77	0.31
4	-1.10	0.00	1.10	1.00
5	0.35	0.59	0.96	1.08

Table 2. Item Parameters of the Essay Test by Empiric

Item Number	b 1	b 2	b 3	a
1	3.37	-3.60	3.13	5.24
2	-1.10	0.00	1.10	1.00
3	-0.11	3.85	26.27	0.96
4	-1.10	0.00	1.10	1.00
5	3.18	-4.13	4.10	6.46

The estimation of testees' ability in the empirical study in the testlet in the interval between 1.33 and 1:33. While the estimation of testees' ability in the empirical study in the essay test zn the interval between -1.33 and 0.14. The estimation results of testees' ability in the empirical study is used to generate the true paramater with the sample sizes of 500.

The true zrom the testlets and the essay test is used to generate the simulation data with test length by 5 items and sample sizes to 500. Each set was *replicated* 25. Correlation between the testees' ability in simulation and true score for each replication in the essay test and testlets are shown in Table 3. RMSEA value for the testees' ability in the testlet and essay test are shown in Table 4.

As shown in Table 1 and 2, the estimation results of testees' ability in the empirical study show that the use of testlets are overestimate than essay test. It is seen that the testees' ability estimation on the testlet ranged from -1.33 to 1.33 while on the essay test, the testees' ability estimation ranged from -1.33 to 0.14. The distribution of testees' ability in the essay test is 40% in -1.33. This means empirically, testees'ability appear higher on the testlets than on the essay test thus testlets is considered easier than essay test.

Table 3. Correlation between the testees' ability in simulation and true score for each replication in essay test and testlets

Number	replication	Testlets	Essay test
1	1	0.420529	0.656814
2	2	0.465972	0.617115
3	3	0.408278	0.593128
4	4	0.452014	0.661796
5	5	0.472515	0.643351
6	6	0.442672	0.642035
7	7	0.41369	0.616228
8	8	0.38573	0.590539
9	9	0.471735	0.644028
10	10	0.320372	0.642261
11	11	0.416838	0.600648
12	12	0.431693	0.627185
13	13	0.427089	0.648804
14	14	0.376792	0.627232
15	15	0.400892	0.656649
16	16	0.442837	0.633983
17	17	0.451999	0.657832
18	18	0.428893	0.638573
19	19	0.455322	0.631792
20	20	0.406328	0.604489
21	21	0.392176	0.649054
22	22	0.428585	0.627664
23	23	0.397294	0.634545
24	24	0.45824	0.640807
25	25	0.42715	0.650658
	mean	0.423825	0.633488

Table 4. Value of RMSEA of Testees' Ability simulation with the testlet and the essay test

Number	replication	testlets	essay test
1	1	0.499153	0.153767
2	2	0.47228	0.16573
3	3	0.496778	0.160405
4	4	0.479911	0.154997
5	5	0.464823	0.158881
6	6	0.4824	0.152047
7	7	0.491672	0.166123
8	8	0.510259	0.16715
9	9	0.471907	0.160146
10	10	0.539345	0.154061
11	11	0.500709	0.165557
12	12	0.491175	0.161533
13	13	0.490226	0.162092
14	14	0.511503	0.158744
15	15	0.503686	0.156523
16	16	0.487045	0.158521
17	17	0.481408	0.151148
18	18	0.49154	0.15751
19	19	0.477035	0.155428
20	20	0.500337	0.164577
21	21	0.505569	0.158753
22	22	0.487836	0.162303
23	23	0.509775	0.154177
24	24	0.476072	0.152913
25	25	0.488296	0.156561
	mean	0.49243	0.158786

As shown in Table 1 and 2 that the item parameter estimates to the essay test is much more difficult than the testlets. It can conclude that the empirical estimation of item parameter with essay test is irrational until one reaches 26,27. However, these results still credible because the test is made in criteria reference test so that item parameter can have extreme values.

Distribution of true score is still relatively similar to the results of empirical parameter that is to testlets from about -1.33 to 1.33 and in essay test from about -1.33 to 0.14. The mean of true score on testlet is 0.0289 with a standard deviation is 0.03, while the mean of true score in the essay test is -0.587 with a standard deviation is 0.01. Table 3 shows that the correlation coefficient between true scores and parameters of the simulation studies, the essay test tends to have values which are higher than testlets. The mean of correlation coefficient of true score and simulation parameter on the essay test is 0.633 while the testlets is 0.423. This shows that in terms of accuracy, more precise in an essay test than testlet on estimating the testees'ability.

Table 4 shows that the MSE index of 25 replication in simulation studies, the essay test tends to have lower than testlet zThe mean of MSE on the essay test of 0.15 while in the testlets is 0.49. This shows that in terms of precision, more precise in an essay test than testlet on estimating the testees'ability.

From the analysis it can be concluded that the essay test more precise on estimating the ability of testees'ability. The results are consistent with several previous studies. Purwo Susongko (2010) proved that essay test is more effective for chemical achievement test than testlets with the application of Graded Response Model. The study was conducted empirical and simulation using the item information functions as a criterion of effectiveness.

Through his studies, Zidner (1987: 607) conclude that the essay test requires a high ability to organize a response, requires the ability to recall the material, requires integrative knowledge and the ability to write well. In the multiple choice test is not found anything like that, because the testee just choose the option that has been prepared. If a testee answered correctly for the same items on the multiple-choice test, it is difficult to presume that the option based on the results of complex thinking.

Kuechler and Simkin (2003: 394) through theirs study concluded that in the multiple choice test, students have a chance to guess the correct answer is greater than in the essay test . Shepard, (2008: 604), through a study conducted by the National Mathematics Advisory Panel of more than 15 studies, concluded that: (1) an error when the multiple choice test and essay test used to measure the competence of the same, (2) the essay test is used to measure the higher ability of students and (3) the essay test has more information than multiple choice test .

IV. CONCLUSIONS AND RECOMMENDATION

A. Conclusions

1. The accuracy of the estimation of testees' ability of mathematics achievement essay test is higher than testlets,
2. The precision of the estimation of testees' ability of mathematics achievement essay test is higher than testlets

B. Suggestion

1. It needs further research about differences in the accuracy of the essay test and testlets in estimating the testees' ability by considering other variables such as the length of the test, the sample size and the number of replication.
2. Need more in-depth research about differences in the accuracy of the essay test and testlets in estimating the testees' ability with item response theory modeling of other types such as the modeling GPCM, PCM and other models politomos models.
3. Keep a study similar to the field of educational measurement other areas such as in the field of language studies, science and social sciences.
4. In practice, a essay test should be more widely used than the form testlet in educational measurement.

REFERENCES

[1] Ferrara, S, Huynh, H, & Baghi, H. (1997). Contextual characteristics of locally dependent open-ended item cluster in a large-scale performance assessment. Applied Measurement in Education, 10(2),123-144

[2] Fischer, A. R & Tokar, D. M. (1998). Validity and construct contamination of the racial identity attitude scale-long form. Journal of Counceling Psychology, 45(2),212-224.

[3] Hamilton, Clayon B; Maly, Monica R; et al (2015), Health and Quality of Life Outcomes 13 Validation of the Questionnaire to Identify Knee Symptoms (QuIKS) using Rasch analysis http://search.proquest.com/docview/1780120342?accountid=62691

[4] Heri Retnowati (2006). Stabilitas estimasi Parameter pada Regresi Logistik (suatu penerapan pada pengukuran). http://eprints.uny.ac.id/7246/1/PM-9%20-%20Heri%20Retnawati.pdf

[5] Lee, G.et al. (2001). Comparisons of dichotomous and polythomous item respons models in equiting scores from test composed of teslets. Applied Psychological Measurement , 25 (4), 357-372.

[6] Purwo Susongko, (2009). Studi Komparasi Kualitas Butir Tes Prestasi Belajar Matematika Pada Tes Bentuk Uraian Dan Bentuk Testlet .Penelitian Mandiri. tidak diterbitkan

[7] Purwo Susongko, (2009).Perbandingan Keefektifan Bentuk Tes Uraian dan Testlet Dengan Penerapan Graded Response Model (GRM), Disertasi, Tidak diterbitkan, Pascasarjana UNY Yogyakarta, 2009

[8] Purwo Susongko, (2014). Perbedaan ketepatan estimasi tingkat kesukaran butir tes pilihan ganda pada penskoran koreksi dan penskoran konvensional dengan penerapan model Rasch. Makalah tidak diterbitkan. Dipresentasikan pada Seminar nasional Matematika dan pendidikan Matematika, USD Yogyakarta, 13 September 2014.

[9] Risky Setiawan (2014) Analisis Dan Simulasi Dengan Program Win-Gen(Strategi Dalam Mengkonstruk Instrumen Soal). e-journal.ikip-veteran.ac.id/index.php/pawiyatan/.../8

[10] Shepard, L.A.(2008). Commentary on the national mathematics advisory panel recommendations on assessment. Educational Researcher, 37(9),602-609.

[11] Van der Linden, W. J & Hambleton, R. K. (1997). Handbook of modern item response theory. New York ,NY:Spring-Verlag, Inc.

[12] Verhelst, N. D & Verstralen, H. H. F. M. (1997). Modeling sums of binary responses by the partial credit model. Measurement and Research DepartementReports, 1-18

[13] Wainer, H. (1995). Precision and differential item functioning an a teslets based test : The 1991 law school admissions test as an example. Applied Measurement in Education, 8, 157-186.

[14] Wainer, H, Sireci, S. G & Thissen, D. (1991). Differential testlets functioning: Definition and detecting . Journal of Education Measurement, 24, 185-201.

[15] Wainer, H & Wang, X. (2000). Using a new statistical model for teslets to score TOEFL.Journal of Educational Measurement,. 37, 203-220

[16] Zedner, M.(1987). Essay versus multiple-choice type classroom exams: the student perspective. Journal of Educational Research, 80(6),352-358.

[17] Zenisky, A., Hambleton, R & Sireci, S. G (2002). Identification an evaluation of local item dependencies in the medical college admission test, Journal of Educational Measurement, 39 (4), 291-309