Characteristic Analysis of Essay Test Instruments for Measuring Higher-Order Thinking Skills

E Susilaningsih[1], D L Setyowati[2], A M Diputera[3]
[1,2,3]Graduate School, Universitas Negeri Semarang, Indonesia
[1]Corresponding email: endang.arkan@gmail.com

Abstract

This study aims to test the reliability and items characteristics of the essay test instrument to measure the higher-order thinking skills of social science lessons of junior high school grade 8. This research is useful to know the characteristics of the instrument items used directly by teachers for measurement activities in learning. This research method using quantitative research techniques. The sample used for the test used 105 students taken at random. The result of the research shows the estimation of instrument reliability coefficient of 0.819. The grain characteristics test of the difficulty level parameter for 15 items consists of eight items in the Easy category, 6 items in Medium category and one item in difficult category. Grain characteristics test for different power parameters found three items that are not able to distinguish the ability of students. The conclusion of this research is reliable instrument to measure the ability of high-level thinking and 12 items can be able to distinguish students' thinking ability. Teachers can use instruments consisting of 12 items to measure high-order thinking skills.

Keywords: instrument, essay test, higher order thinking skills.

1. Introduction

Teachers in learning must act objectively and not discriminate against the students. The teacher designed the learning activities instinctively to help the students and know to meet the needs of each student. The teacher knows where to start and help according to his ability (Richburg & Nelson (1998), Fatmasuci, (2017)). Teachers need to include learning to help students train problem solving.

The teacher knows about problem solving and is not new (Hidayah, Suyitno, & Junaedi, 2014). Problem solving considered as an objective oriented process that uses integrated higher-order thinking skills, such as generating ideas, creating interpretations and judgments, and using the complexity of circumstances (Kirkwood, 2000, p 511, Sucipto, 2017). Measurement of troubleshooting capabilities can use contextual troubleshooting questions (Samo, 2017).

Schraw in Kusuma, Rosidin, Abdurrahman, & Suyatna (2017, pp. 26-32) classifies the thinking based on the taxonomy developed by Bloom into two categories. The category consists of a Lower Order Thinking Skill that consists of knowledge, understanding and application. Higher Order Thinking Skill consists of Analyse, Synthetic, and Evaluation. (Krathwohl & Anderson, 2010, p 215) revised Bloom's taxonomy consisting of Remember, Understand, Apply, Analyse, Evaluate, and Create. Problems that develop high-order thinking skills have a relatively low percentage (Juhanda, 2016).

Students in Indonesia as much as 1% have advanced thinking skills and 78% have low-level thinking skills from Taiwan, South Korea, Singapore, Hongkong and Japan who have high-level thinking skills above 40% (Nusarastriya, 2013, p 24, Kurniati, Harimukti, & Jamil, 2016). Teachers can measure students' high-order thinking skills using to test techniques.

The test is a procedure that contains sequential steps, contains a sample of behaviour and measures behaviour. Essay tests tend to have higher information functionality than multiple choices (Sasongko, 2010, La Fave, 1966). Topics taught require feedback from students rather than just choosing answers then it should be developed item polytomous (Ridlo, 2011, p.41). The item in the test requires students to show what they will find out by answering questions (Azwar, 2010, p.3, Hambelton & Rogers, 2000, p.4).

The ability to write an essay test may link scores of writing skills in admissions rather than student placement tests (Goodwin, 2016). Essay tests developed to cover the domain of clinical judgment to provide information (Day et al., 1990).

Students must have higher-order thinking skills in the form of critical thinking skills in order to be ready to face the changing circumstances in the learning process (Arafat, Ridlo, & Priyono, 2012, p.48). Higher-order thinking skills is important aspects to develop in learning (Susanto & Retnawati, 2016). However, it seems that seventh and eighth grade science teachers are individualistic and diverse in reference to teaching techniques (Lawrenz, 1990).

Preparation of items on higher order thinking using problem-based learning should take account of educational goals. Teachers should plan and design problems to meet the objectives to be achieved (Weiss, 2003). Project-based learning proven to help students become collaborators, develop thinking skills, share ideas and discuss ideas, find and analyse information on multiple sources and create multimedia presentations (Susanawati, Diantoro & Yulianti, 2014, Faizah et al., 2015, Nuswowati, Susilaningsih, Ramlawati, & Kadarwati, 2017).

Barnett & Francis (2012) conducted a study to test whether quizzes containing high-order thinking questions related to critical thinking and performance tests when used simultaneously. The results show that critical thinking increases equally in all sections. The sections that receive the higher-order thinking quiz done significantly better than the other two sections in the multiple choice and essay sections.

The researcher revised the item and scoring guidance for a wider scope test. This study aims to examine the characteristics of the grains of standard essay test instruments to measure the thinking ability of the high level of social science subjects of 8th grade high school that have been developed. The benefit of this research is to describe the characteristics of the grains so that the instrument is ready for use by the teacher.

2. Method

This research uses quantitative research techniques. The study conducted in junior high school 4 Gununghalu Regency West Bandung. Samples taken from the population at random. Researcher used error rate 5% and trust to population equal to 95% so that sample used in this research is 105 students on class 8.

Data collection technique in this research is using test technique. The test instrument used is a test instrument developed by researchers (Diputera, Setyowati, & Susilaningsih, 2018). The results were analysed quantitatively using IBM SPSS version 24 and Microsoft Excel software. Analysis includes item validity (r_{xy}), difficulty index (P), discrimination index (D) and Instrument Reliability (r_{11}).

3. Result and Discussions

An empirical test using 105 students resulted in a score is correlated with its total score to determine the validity of the item (r_{xy}). The result of items validity test result from 15 items; 13 item is valid and two is not valid. Item declared invalid because it has a value <0.3. The test results of item validity seen in Table 1.

Table 1. Results of Item Validity Analysis

Item	r_{xy}	Category	Item	r_{xy}	Category
1	0.73	Valid	9	0.73	Valid
2	0.10	Invalid	10	0.56	Valid
3	0.67	Valid	11	0.81	Valid
4	0.87	Valid	12	0.72	Valid
5	0.04	Invalid	13	0.62	Valid
6	0.58	Valid	14	0.33	Valid
7	0.69	Valid	15	0.64	Valid
8	0.52	Valid			

The researcher analysed the difficulty of large-scale test results. The results of the analysis show that 15 grains have varied degrees of difficulty in *the Easy*, *Moderate*, and *Difficult* categories seen in Table 2. The investigators analysed the degree of difficulty of large-scale test results. The results showed that eight grains had difficulty in the *Easy* category. Six items have difficulty level in *Moderate* category. One item has *Difficulty* category difficulty index (P).

Table 2. Analysis of difficulty index

Item	P	Category	Item	P	Category
1	0.76	Easy	9	0.64	Moderate
2	0.43	Moderate	10	0.76	Easy
3	0.75	Easy	11	0.75	Easy
4	0.50	Moderate	12	0.61	Moderate
5	0.38	Moderate	13	0.75	Easy
6	0.75	Easy	14	0.27	Difficult
7	0.75	Easy	15	0.75	Easy
8	0.48	Moderate			

The result of the analysis of the discrimination of the test (D) of the large scale test of standardized test instrument design to measure the high grade thinking grade in the class 8 IPS indicates that from 15 items there are 8 items *accepted*, 2 items *received need to be fixed*, 2 items are *fixed* and 3 items *not used*. The researcher discarded point 2, 5, and 14. The researcher discarded three items that did not have the ability to distinguish because each indicator of achievement still represented. Items arranged based on the 12 items received to see Table 3.

Table 3. Results of Discriminant Index Analysis

Item	D	Category	Item	D	Category
1	0.28	problem fixed	9	0.50	Accepted
2	-0.34	not used	10	0.51	Accepted
3	0.44	accepted	11	0.37	received needs to be fixed
4	0.59	accepted	12	0.46	Accepted
5	-0,35	not used	13	0.50	Accepted
6	0.53	accepted	14	-0.01	not used
7	0.36	received needs to be fixed	15	0.44	Accepted
8	0.23	problem fixed			

The test results in the standard test instrument reliability test (r_{11}) to measure the thinking ability of the high level of social science subjects of junior high school grade 8 shows an estimated value greater than 0.7. Instrument reliability test performed using 15 items of question and instrument reliability test using 12 items that have eliminated three items that were not able to differentiate students' ability to see Table 4.

Table 4. Data Analysis Reliability

	Cronbach's Alpha	N of items
Large Scale	0.819	15
Final test	0.918	12

Table 4 shows that the coefficient of large-scale test reliability shows the reliability coefficient (r_{11}) of large-scale test of 0.819. The final test of 12 grains declared accepted on large-scale test showed a reliability of 0.918.

Twelve (12) items declared ready to use based on item validity analysis (r_{xy}), difficulty index analysis (P), discriminant index analysis (D) and reliability analysis (r_{11}). The researchers eliminated three items that were otherwise incapable of distinguishing students' abilities. The 12 items that are ready for use.

The essay test instrument to measure the thinking ability of high level of social science subject of junior high school grade 8, which compiled then analysed the characteristics and reliability (r_{11}) in this research. Grain characteristics analysis consists of item validity analysis (r_{xy}), discriminant index analysis (P), difficulty index analysis (D) and instrument reliability estimation (r_{11}).

The validity of the item provides an overview of the conformity of the item by correlating the acquisition score of its total score. The result of item validity test (r_{xy}) known from 15 items tested to produce 13 items declared valid. Items are valid in accordance with the requirement that the item must be ≥0.3. Items that are below 0.3 then declared invalid.

The analysis of the item difficulty index (P) yields a description of the difficulties index that students face to answer questions. The difficulty index (P) of 15 items of problem found eight items in the easy category, 6 items are in the medium category and 1 item is in the difficult category. The value of the difficulty index (P) provides information that out of 15 points of question has a considerable degree of difficulty.

Discriminant index analysis (D) results in an item's ability to differentiate students' abilities. Students who classified clever should be able to answer questions and students who classified as less intelligent should not be able to answer. However, if it is believed the question is not able to distinguish the ability of students. Discriminant index (D) of 15 items found eight items on the category received, 2 items received categories need to be improved, 2 items of the category improved and 3 item categories removed.

The items declared invalid are 2 and 5. Item 2 obtains a validity coefficient (r_{xy}) of 0.10 and item 5 obtains a validity coefficient (r_{xy}) of 0.04. Items 2 and 5 are well below standard, so they are not valid items. Items have no correlation between the earning score and the total score. Item 2 contains the question "Compare the process and reaction to the Indonesian nation upon the arrival of the three Western nations?". Questions contain the ability to evaluate. Students have not been able to compare the process and reaction to the Indonesian nation upon the arrival of the western nation. The question is quite difficult because it must evaluate the arrival of three nations. Students are difficult to analyse the entry process of the three nations by mentioning the characters, time and process. Students are unable to evaluate the Indonesian response because the material does not contain explicitly the reaction to the Indonesian nation to the arrival of the western nation.

Item 2 is not valid item with the support of the difficulty index (P) and discriminant index (D). The analysis of the difficulty index (P) for item 2 gets the moderate category seen in Table 2. However, being on the numbers is almost close to difficult. Students believed to be quite difficult to analyse and write it in a table. The discriminant index analysis (D) of item 2 also results in the category "not used" to refer to Table 3, as it is unable to distinguish the student's ability.

Item 5 contains the question "Compare the objectives and rules of implementation of forced to labour policies, land to rent systems, and forced cultivation systems during colonial times?". Questions contain the ability to evaluate. Students have not been able to compare the objectives and policy execution of various western peoples. The question is quite difficult because it must evaluate the policies of different nations. Students find it difficult to evaluate the objectives and rules of conduct. Students are unable to decipher because the purpose of the policy that does not explicitly explain and the rules are quite a lot. Students have not been able to evaluate to compare the three policies.

Item 5 does not express a valid item with the support of an index of difficulty and a discriminant index. The analysis of the level of difficulty for item 5 gets Category Referring to Table 2, students become quite difficult to analyse and write it in table form. Difficulty index analysis (P) of point 5 also results in the category "not used" to refer to Table 3, as it is unable to distinguish the student's ability.

Item 14 declared valid item based on the validity test of the item. However, the level of difficulty and differentiation produce bad value. Item 14 obtains the difficulty level in the difficult category and the different matter of getting the category of matter removed. Students find it difficult to answer and the items are not able to differentiate students' ability. Students are unable to decipher the form and relationship of the struggle for the Japanese occupation.

Based on the result of item validity analysis (r_{xy}), difficulty index (P) and discriminant index (D) then only 12 items are declared ready and proper for teacher use. Twelve items declared valid, had varying degrees of difficulty and were able to differentiate students' ability well.

The reliability estimation (r_{11}) of the instrument in this study divided into 2 test sections. The first section tests the reliability for 15 items and section 2 tests the reliability of 12 items that declared ready and feasible. The reliability estimate uses the standard 0.7, so the instrument reliability coefficient must be greater than or equal to 0.7.

The reliability coefficient (r_{11}) for 15 items obtained a value of 0.819. Reliability coefficient (r_{11}) of 12 items got a higher value of 0.918. These results provide information that based on empirical test of the instrument using 12 items that declared ready and feasible to use have consistency and trust assessment is very high.

3. Conclusion

The essay test instrument to measure higher-order thinking skills has tested the characteristics of producing 12 items that are ready and feasible to use. Two items declared invalid because they have no correlation between their total score supported by the result of difficulty index analysis and discriminant index. One item declared valid, but has a difficulty level in the Sukar category and is unable to distinguish students' abilities. The grain difficulty analysis resulted in eight items in the Easy category, 6 medium category items and 1 difficult category grains. The 15 items tested for different powers resulted in 12 items that were able to differentiate students' abilities.

Estimation of reliability of 15 items is reliable and 12 items that are ready and feasible to use are considered reliable with very high category. Teachers should use an essay test instrument to measure high-order thinking skills in measuring. The instrument is valid and reliable, so it can give maximum measurement result.

4. References

Arafah, S. F., Ridlo, S., & Priyono, B. (2012). Pengembangan LKS Berpikir Kritis Pada materi Animalia. *Unnes Journal of Biology Education, 1*(1), 47–53.

Azwar, S. (2010). *Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar*. Yogyakarta: Pustaka Pelajar.

Barnett, J. E., & Francis, A. L. (2012). Using higher order thinking questions to foster critical thinking: a classroom study. *Educational Psychology*, *32*(2), 201–211. https://doi.org/10.1080/01443410.2011.638619

Day, S. C., Norcini, J. J., Diserens, D., Cebul, R. D., Schwartz, J. S., Beck, L. H., … Elstein, A.(1990). The validity of an essay test of clinical judgment. *Academic Medicine*, *65*(9), S39-40. https://doi.org/10.1097/00001888-199009000-00034

Diputera, A. M., Setyowati, D. L., & Susilaningsih, E. (2018). Higher-Order Thinking Skills of Junior High School Students. *The Online Journal of New Horizons in Education*, *8*(3), 61–67.

Faizah, U., Prastiwi, M. S., Subekti, N., Setyowati, D. L., Rachmadiarti, F., & Kuncjoro, S. (2015). Teaching Materials Model-Based Problem Based Learning (PBL) to Habituate Students Conservation. In *Procceding ICCBL 2015* (pp. 1–3). Semarang: International Conference on Conservation for Better LIfe.

Fatmasuci, F. W. (2017). Pengembangan perangkat pembelajaran berbasis masalah berorientasi pada kemampuan komunikasi dan prestasi belajar matematika siswa SMP. *Jurnal Riset Pendidikan Matematika*, *4*(1), 32. https://doi.org/10.21831/jrpm.v4i1.11325

Goodwin, S. (2016). A Many-Facet Rasch analysis comparing essay rater behavior on an academic English reading/writing test used for two purposes. *Assessing Writing*, *30*, 21–31. https://doi.org/10.1016/j.asw.2016.07.004

Hambelton, R. K., & Rogers, H. J. (2000). *Advance in Criterion-Referenced Measurement*. New York: Springer Sciences+Business Media.

Hidayah, N., Suyitno, H., & Junaedi, I. (2014). Analisis Kemampuan Guru Matematika SMP dalam Membuat Soal-Soal Pemecahan Masalah. *Unnes Journal of Mathematics Education Research*, *3*(1). Retrieved from https://journal.unnes.ac.id/sju/index.php/ujmer/article/view/7017

Juhanda. (2016). Analisis Soal Jenjang Kognitif Taksonomi Bloom Revisi Pada Buku Sekolah Elektronik (BSE) Biologi SMA. *Jurnal Pengajaran Matematika Dan Ilmu Pengetahuan Alam*, *21*(1). Retrieved from http://journal.fpmipa.upi.edu/index.php/jpmipa/article/view/657

Kirkwood, M. (2000). Infusing higher-order thinking and learning to learn into content instruction: A case study of secondary computing studies in Scotland. *Journal of Curriculum Studies*, *32*(4), 509–535. https://doi.org/10.1080/00220270050033600

Krathwohl, D. R., & Anderson, L. W. (2010). Merlin C. Wittrock and the revision of bloom’s taxonomy. *Educational Psychologist*, *45*(1), 64–65. https://doi.org/10.1080/00461520903433562

Kurniati, D., Harimukti, R., & Jamil, N. A. (2016). The Higher Order Thinking Skills of Junior High School Students at Jember District in Solving PISA Standar-Based Test Item. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 142. https://doi.org/10.21831/pep.v20i2.8058

Kusuma, M. D., Rosidin, U., Abdurrahman, A., & Suyatna, A. (2017). The Development of Higher Order Thinking Skill (Hots) Instrument Assessment In Physics Study. *IOSR Journal of Research & Method in Education (IOSRJRME)*, *7*(1), 26–32. https://doi.org/10.9790/7388-0701052632

La Fave, L. (1966). Essay vs. multiple-choice: Which test is preferable? *Psychology in the Schools*, *3*(1), 65–69. https://doi.org/10.1002/1520-6807(196601)3:1<65::AID-PITS2310030117>3.0.CO;2-Y

Lawrenz, F. (1990). Science teaching techniques associated with higher-order thinking skills. *Journal of Research in Science Teaching*, *27*(9), 835–847. https://doi.org/10.1002/tea.3660270904

Nusarastriya, Y. H. (2013). Permasalahan Dan Tantangan Guru PKn Menghadapi Perubahan Kurikulum (2013). *Satya Widya*, *29*(1), 23. https://doi.org/10.24246/j.sw.2013.v29.i1.p23-29

Nuswowati, M., Susilaningsih, E., Ramlawati, R., & Kadarwati, S. (2017). Implementation of Problem-Based Learning with Green Chemistry Vision to Improve Creative Thinking Skill and Students' Creative Actions. *Jurnal Pendidikan IPA Indonesia*, *6*(2), 221. https://doi.org/10.15294/jpii.v6i2.9467

Richburg, R. W., & Nelson, B. J. (1998). Integrating Content Standards and Higher-Order Thinking: A Geography Lesson Plan. *The Social Studies*, *89*(2), 85–90. https://doi.org/10.1080/00377999809599830

Ridlo, S. (2011). Pengembangan Tes Pengetahuan Praktikum Biologi Berdasarkan GRM dan GPCM. *Jurnal Pendidikan Matematika Dan Sains*, *1*(XVI), 41–49.

Samo, D. D. (2017). Kemampuan pemecahan masalah matematika mahasiswa tahun pertama dalam memecahkan masalah geometri konteks budaya. *Jurnal Riset Pendidikan Matematika*, *4*(2), 141. https://doi.org/10.21831/jrpm.v4i2.13470

Sasongko, P. (2010). Comparison of The Effectiviness of The Essay Test And Testlets Through The Graded Response Model (GRM) Application. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *14*(2). Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/1082/865

Sucipto, S. (2017). Pengembangan Ketrampilan Berpikir Tingkat Tinggi dengan Menggunakan Strategi Metakognitif Model Pembelajaran Problem Based Learning. *Jurnal Pendidikan (Teori Dan Praktik)*, *2*(1), 77. https://doi.org/10.26740/jp.v2n1.p77-85

Susanawati, E., Diantoro, M., & Yulianti, L. (2014). Pengaruh Strategi Projectbased Learning Dengan Thinkquest Terhadap Kemampuan Berpikir Kritis Fisika Siswa SMA Negeri 1 Kraksaan. *Jurnal Pengajaran Matematika Dan Ilmu Pengetahuan Alam*, *18*(2), 207. https://doi.org/10.18269/jpmipa.v18i2.51

Susanto, E., & Retnawati, H. (2016). Perangkat pembelajaran matematika bercirikan PBL untuk mengembangkan HOTS siswa SMA. *Jurnal Riset Pendidikan Matematika*, *3*(2), 189. https://doi.org/10.21831/jrpm.v3i2.10631

Weiss, R. E. (2003). Designing Problems to Promote Higher-Order Thinking. *New Directions for Teaching and Learning*, *2003*(95), 25–31. https://doi.org/10.1002/tl.109