JERE 8 (2) (2019) 141 - 147

Journal of Educational Research and Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Validity and Reliability of Higher Order Thinking Test Assessment (HOTS) in Mathematics Learning At Seventh Grade Based on The Expert Study

Resvia Subay ✉**, Kartono, Sulhadi**

Universitas Negeri Semarang, Indonesia

Article Info

Article history:
Received 25 January 2018
Approved 12 March 2019
Published 23 August 2019

Keywords:
Validity, Reliability, HOTS Test Assessment

Abstract

This paper is a part of research and development that develops Higher Order Thinking Skill test assessment in mathematics learning in seventh grade. The developed test assessment is in the form of essay questions, consisting of 10 question items. In a research and development study, one of the required steps before conducting a trial run test is an expert study. The expert study is carried out to find out the validity and reliability of a developed instrument based on experts' judgment. The developed test assessment in this research is based on three experts' judgments. They are a mathematician, an evaluation expert, and a practitioner (mathematics teacher). Then, the scores taken from the expert study were then analyzed by using the Aiken formula. Meanwhile, the reliability was analyzed by using Ebel. The validity score of 10 question items shows each of them has a validity score ≥0,67toward three aspects. The aspects are content, construct, and language. The aspects consist of 14 principles. Meanwhile, the reliability test has a score of 0,96. Thus, the developed test assessment, based on the expert study, is valid and reliable.

✉Correspondence Address:
Kampus Pascasarjana UNNES Jl Kelud Utara 3 Sampangan
Semarang, Indonesia
E-mail: resviasubay@gmail.com

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Mathematics is a lesson taught for school students. It has educational roles to achieve educational goals as mandated by the Constitution (Pulungan, D. A. 2014). Mathematics is basic science. By learning mathematics, critical, logical, and systematic skills could be trained. However, it is important to know that mathematics roles do not only cover those matters but mathematics has roles in other fields, for example, physics, economy, and biology. Thus, in learning mathematics, it is important to be directed in higher thinking order skills (Sa'idah, N., Yulistianti, H. D., & Megawati, E. 2019).

The 2013 curriculum is the present-developed curriculum that demands students in learning mathematics could reach higher-order thinking skills (Wanda, V. N., Yusmin, E., & Nursangaji, 2019). Ferita, R. A., & Fitria, M (2019) argue that student-critical thinking skill improvement, in the present time, is the focus attention of the government. It could be seen by the implementation of the curriculum in which higher-order thinking skill of students becomes its focus. Furthermore, higher-order thinking skill gets special attention from one of International study institution called TIMMS (*The Trends For International Mathematics and Science Study*). The institution reviews student cognitive skills in the mathematics and science field (Fitriani, D., Suryana, Y., & Hamdu, G., 2018).

Higher-order thinking skills cover skills to analyze, evaluate, and create (Ferita, R. A., & Fitria, M., 2019). Higher-order thinking skills in mathematics, known as MathHOT, are skills to analyze, evaluate, and create in the mathematics field. These skills could be performed in completing mathematics problems by analyzing, evaluating, and creating (Hikmah, H., & Amin, N. 2019).

A study conducted by Imanudin, T.n.F. (2015) concluded that JHS/Islamic JHS students' books in the seventh grade with 2013 curriculum, published by Ministry of Culture and Education in 2014, the first-semester revision edition, consisted of 74 questions, a 2.72% C4 question percentage, a 2.72% C5 question percentage, and a 0% C6 question percentage. Furthermore, Budiman, A., & Jailani, J (2014) stated the teachers' hindrances were lack of skills to develop HOTS assessment instruments. The previous arguments are also in line with Pulungan's arguments (2014) about the incompatibility of the curriculum demands and the availabilities of measuring tools in the form of test instruments. Budiman, A. & Jailani, J (2014) argue that the given questions by teachers influence students' thinking skill developments. Higher-order thinking skills could be trained by giving questions that trigger students to think analytically, creatively, and in an evaluative manner.

Identifying how far students' higher-order thinking skills are could be done by conducting an assessment. Assessment is an applicable test to train students' thinking skills. It also influences in determining the students' thinking skills. (Suhaesti Julianingsih, S. J., Undang Rosidin, U. R., & Ismu Wahyudi, I. W. 2017). Furhtermore, Dhema, M. (2019) states that the measuring instruments in an assessment should have high-reliability quality criteria to be applied in measuring students' competences. Besides that, Ferita, R. A., & Fitria, M. (2019) adds that the excellent test instruments to be applied as measuring instruments should be valid and reliable.

Validity is an important matter in developing both test and nontest instruments (Mardapi Dhemari, 2016:33). Furthermore, Sugiyono (2015:179) states that valid instruments must have both internal and external validities. What is meant by the internal validity of an instrument is rationale validity. Internal validity must consist of construct and content validities. However, Ramalis, T.R., & Purwana, U (2018) state that content validity consists of material, construct, and language aspects in an instrument. This validity evidence could be an accuracy analysis of the test content logically or empirically to create score-test interpretations. The test or instrument validity evidence should be conducted by experts of the expertises or fields in the measured field and experts of measuring field (Mardapi Djemari, 2016:33-34).

Besides validity, reliability needs to be analyzed carefully. Nugroho, B. S., Djuniadi, D., & Rusilowati, A. (2016) explain that after finding out the validity results based on the experts' judgment, the next step is to calculate the instrument reliability through consistency agreement among the raters. It is due to reliability could show how far the measuring results of the instrument could be trusted (Munadi, S., 2010).

Based on the explanation, in this paper, the validity and reliability of *Higher Order Thinking Skills* (HOTS) test assessment development in learning mathematics at seventh grade with Rasch Model, based on expert study will be explained.

METHOD

Analyzing validity and reliability based on the expert study is a part of this *Higher Order Thinking Skills* (HOTS) test assessment development in learning mathematics four seventh grade with the Rasch Model. The developed test questions are 10 items of mathematics essay for seventh graders in the odd semester. The questions would be reviewed by three experts. They were a mathematician, an evaluation expert, and a practitioner (a mathematics teacher). Analyzing validity scores based on the expert study was done by using the Aiken formula assisted by Microsoft Excel.

Aiken formulates Aiken's formula to calculate the *content-validity coefficient* that is based on judgment results of expert panels with n persons to a certain system. It is based on how far the items represent the measured construct (Hendryadi, H., 2017).

Retnawati Heri (2015:18-19) states that validity is determined by expert agreement. To find out this agreement, validity indexes could be used. One of them is what is proposed by Aiken as follows.

$$V = \frac{\sum s}{n(c-1)}$$

Remark:

$V =$ The index of the rater agreement about item validity.

$s =$ The determined scores of each rater is subtracted by the lowest score in the applied category. $(s = r - I_0)$

$c =$ The numbers of categories that could be selected by the raters.

$n =$ The numbers of the raters.

This index validity is ranged from 0 until 1. The category of this content validity, in which an item could be categorized based on the index, is shown in Table 1.

Table 1. Content validity category.

Scores	Categories
≤ 4	Invalid
$0{,}4 - 0{,}8$	Little bit valid
$> 0{,}8$	Very valid

The reliability of the developed test analysis is analyzed by the Aiken formula. However, an analysis of independent t-test with Two Way ANOVA assisted by SPSS program 20.0 was conducted. The obtained scores were then substituted in the Aiken formula. It is in line with what Azwar (2012) did in A. Nugroho. B. S., Djuniadi, D., & Rusilowati, A. (2016) argue that the instrument reliability is based on the expert agreement. It could be tested by an independent t-test with *Two Way Anova* assisted by SPSS Program 20.0. Then, the analysis was continued by Ebel formula.

Dewi, N. M. A. K., Sugihartini, N., Kesiman, M. W. A., & Sunarya, I. M. G. (2014)state that reliability test among raters is analyzed by using Ebel formula, by looking at the reliability criteria among raters as shown in Table 2.

Table 2. Reliability Criteria among Raters

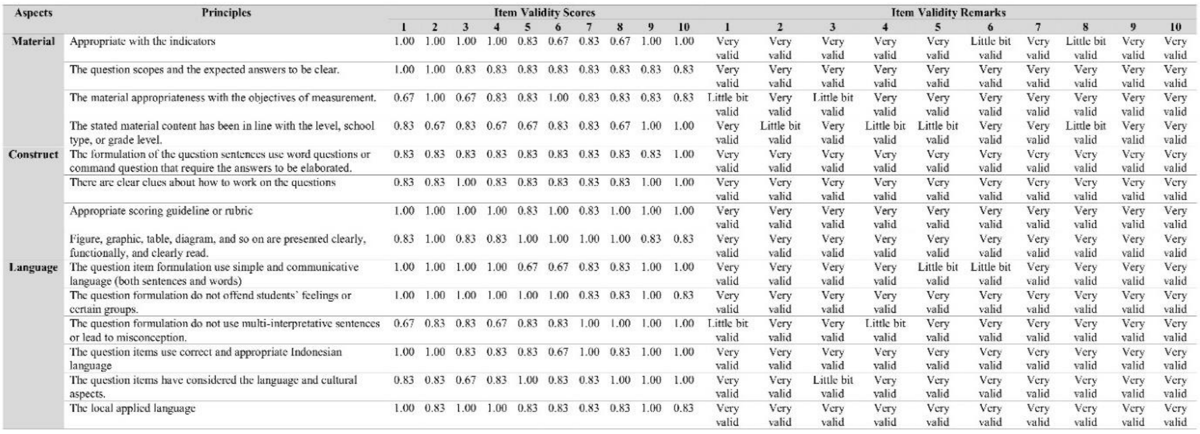
Score Range	Categories
0.00 < r ≤ 0,20	A very low reliability level
0,20 < r ≤ 0,40	Low reliability level
0.40 < r ≤ 0,60	Moderate reliability level
0.60 < r ≤ 0,80	High reliability level
0.80 < r ≤ 1,00	Very high-reliability level

By referring to the previous study conducted by Sujarwanto & Rusilowati (2015), then the Assessment Instrument is stated reliable if the reliability coefficient is ≥ 0,6. Azwar (2000) state that a high-reliability coefficient score could be defined that the given rates by each rater are consistent (reliable) each other.

FINDINGS AND DISCUSSION

The mathematics-essay question items for the seventh grade consist of 10 items. They were reviewed by three experts. They were a mathematician, an evaluation expert, and a practitioner (a mathematics teacher) toward 3 aspects. The aspects are content, construct, and language. Those aspects consist of 14 principles. The review results were then analyzed by applying the Aiken formula and using Microsoft Excel. The analysis results are shown in Table 3

Table 3. The validity scores based on three experts' judgment



Aspects	Principles	Item Validity Scores										Item Validity Remarks									
		1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
Material	Appropriate with the indicators	1.00	1.00	1.00	1.00	0.83	0.67	0.83	0.67	1.00	1.00	Very valid	Very valid	Very valid	Very valid	Very valid	Little bit valid	Very valid	Little bit valid	Very valid	Very valid
	The question scopes and the expected answers to be clear.	1.00	1.00	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The material appropriateness with the objectives of measurement.	0.67	1.00	0.67	0.83	0.83	1.00	0.83	0.83	0.83	0.83	Little bit valid	Very valid	Little bit valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The stated material content has been in line with the level, school type, or grade level.	0.83	0.67	0.83	0.67	0.67	0.83	0.83	0.67	1.00	1.00	Very valid	Little bit valid	Very valid	Little bit valid	Little bit valid	Very valid	Very valid	Little bit valid	Very valid	Very valid
Construct	The formulation of the question sentences use word questions or command question that require the answers to be elaborated.	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.83	1.00	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	There are clear clues about how to work on the questions	0.83	0.83	1.00	0.83	0.83	0.83	0.83	0.83	1.00	1.00	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	Appropriate scoring guideline or rubric	1.00	1.00	1.00	1.00	0.83	1.00	0.83	1.00	1.00	1.00	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	Figure, graphic, table, diagram, and so on are presented clearly, functionally, and clearly read.	0.83	1.00	0.83	0.83	1.00	1.00	1.00	1.00	0.83	0.83	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
Language	The question item formulation use simple and communicative language (both sentences and words)	1.00	1.00	1.00	1.00	0.67	0.67	0.83	0.83	1.00	1.00	Very valid	Very valid	Very valid	Very valid	Little bit valid	Little bit valid	Very valid	Very valid	Very valid	Very valid
	The question formulation do not offend students` feelings or certain groups.	1.00	1.00	1.00	1.00	1.00	1.00	0.83	0.83	1.00	0.83	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The question formulation do not use multi-interpretative sentences or lead to misconception.	0.67	0.83	0.83	0.67	0.83	0.83	1.00	1.00	1.00	1.00	Little bit valid	Very valid	Very valid	Little bit valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The question items use correct and appropriate Indonesian language	1.00	1.00	0.83	0.83	0.83	0.67	1.00	0.83	1.00	1.00	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The question items have considered the language and cultural aspects.	0.83	0.83	0.67	0.83	1.00	0.83	0.83	1.00	1.00	1.00	Very valid	Very valid	Little bit valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid
	The local applied language	1.00	0.83	1.00	1.00	0.83	0.83	0.83	0.83	1.00	0.83	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid	Very valid

Based on the validity scores from the experts, an item numbered 1 has 2 principles that include in moderate validity criteria. It has a validity score of 0,67. The principles are about the content appropriateness to the objective and measurement and the question formulations do not use any word nor sentence causing multi-interpretation or misconception. Meanwhile, the other principle is very valid with validity score ≥ 0,83. Then, the question item number 2 has 1 principle. It is included in moderate validity criteria due to the content material appropriateness to the education levels, type of schools, or the grade level. Meanwhile, the other criteria are included in very valid criteria with validity score≥ 0,83.

The question item numbered 3 has 2 principles that meet moderate validity criteria. They are principles about the content appropriateness to the objective of measurement and the question formulation that has considered the language and cultural aspects. Meanwhile, the other principle in question item numbered 3 is categorized in high validity criteria with validity score ≥ 0,83. The question item numbered 4 has two principles that have moderate validity criteria. The principles are about the content material appropriateness to the educational level, school type, or grade level. The question formulations also do not use any word nor sentences causing multi-interpretation or misconception. Meanwhile, the other criteria of the principles are categorized as high validity with validity score≥ 0,83.

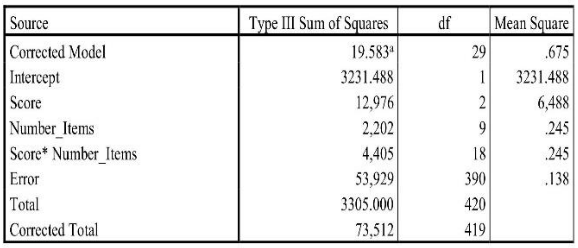
Then, the question item number 5 has 2 principles. They are included in moderate validity criteria due to the content material appropriateness to the education levels, type of schools, or the grade level. The question formulation has considered language and cultural aspects. Meanwhile, the other principles are included in very valid criteria with validity score≥ 0,83. The question item numbered 6 has 3 principles included in moderate validity criteria. They are principles about the question appropriateness to the indicators. The item formulations have implemented simple and communicative language. The question items also used correct and appropriate Indonesian language. Meanwhile, the other principles are included in high validity criteria with validity score≥ 0,83.

The question item numbered 7, in each of its principles has validity score ≥ 0,83. Thus, the question item covers each principle with high validity criteria. The question item numbered 8 has two principles that have validity score 0,67. They are principles about the question appropriateness to the indicators and the questioned-material contents have been in line with the educational level, school type, and grade level. Meanwhile, the other principles are included in high validity criteria with validity score≥ 0,83. The question item numbered 9 and 10 meet each principle. Each of the principles' scores is≥ 0,83. Therefore the question items numbered 9 and 10 have met the high validity criteria.

The validity scores of the developed 10 essay question items, based on expert judgment, had no low validity criteria. Thus, it could be said the developed question items meet the validity criteria based on expert judgment. Thus, the developed test assessment is reliable to use to assess higher-order thinking skill in learning mathematics for the seventh graders in the odd semester.

After obtaining the validity results based on the expert study, the reliability score was analyzed based on the expert study. The results of the ANOVA test assisted by SPSS 20.0 are shown in Table 4.

Table 4. Two Way Anova Test



Source	Type III Sum of Squares	df	Mean Square
Corrected Model	19.583[a]	29	.675
Intercept	3231.488	1	3231.488
Score	12,976	2	6,488
Number_Items	2,202	9	.245
Score* Number_Items	4,405	18	.245
Error	53,929	390	.138
Total	3305.000	420	
Corrected Total	73,512	419	

Then, the obtained score in *Two Way Anova* was substituted in the formula.

$$r_{xx} = \frac{MK_s - MK_{int}}{MK_s}$$

$$r_{xx} = \frac{6{,}488 - 0{,}245}{6{,}488}$$

$$r_{xx} = \frac{6{,}243}{6{,}488} = 0{,}96$$

The obtained reliability score is 0.96. The score shows the reliability criteria among raters have a high-reliability level category. Therefore, the developed instrument, based on the expert judgment, has an opportunity to be trusted to be applied in measuring higher-order thinking skills of the students that the test assessment would have consistent or relatively stable results. Furthermore, due to the high-reliability category score, it means the test assessment instrument has high confidence and there is no difference among the raters.

CONCLUSION

The scoring results of 3 experts obtained validity scores of 10 question items based on three aspects, consisting of 14 principles≥ 0,067. It shows that the developed instrument items meet the validity criteria so the developed test assessment is reliable to assess higher-order thinking skills in learning mathematics in seventh grade.

Meanwhile, the test assessment reliability based on three experts obtained score 0.96. Thus, it means the developed test assessment in learning mathematics for seventh graders has high confidence and it has no differences among raters.

Therefore based on the validity and reliability scores obtained from 3 experts, the developed test assessment could be used to assess HOTS in learning mathematics for seventh graders in the odd semester.

REFERENCE

Alifa, T. F., Ramalis, T. R., & Purwana, U. (2018). Karakteristik Tes Penalaran Ilmiah Siswa Sma Materi Mekanika Berdasarkan Analisis Tes Teori Respon Butir. Jurnal Inovasi dan Pembelajaran Fisika, 5(1), 80-89.

Azwar, S. 2000. Reliabilitas dan Validitas. Yogyakarta: Pustaka Belajar.

Budiman, A., & Jailani, J. (2014). Pengembangan instrumen asesmen Higher Order Thinking Skill (HOTS) pada mata pelajaran matematika SMP kelas VIII semester 1. Jurnal Riset Pendidikan Matematika, 1(2), 139-151.

Dewi, N. M. A. K., Sugihartini, N., Kesiman, M. W. A., & Sunarya, I. M. G. (2014). Pengembangan Instrumen Penilaian Kinerja Penggabungan Gambar 2d Ke Dalam Sajian Multimedia. KARMAPATI (Kumpulan Artikel Mahasiswa Pendidikan Teknik Informatika), 3(1), 73-77.

Dhema, M. (2019). Analisi Instrumen Tes Hasil Belajar Berbasis High Order Thinking Skill (HOTS) Matematika Kelas VII Di SMP Muhammadiyah Waipare. Birunimatika, 4(1).

Ferita, R. A., & Fitria, M. (2019). Pengembangan Instrumen Tes Pilihan Ganda Untuk Mengukur Tingkat Kemampuan Berpikir Matematika Siswa SMA. AKSIOMA: Jurnal Program Studi Pendidikan Matematika, 8(1), 1-10.

Fitriani, D., Suryana, Y., & Hamdu, G. (2018). Pengembangan Instrumen Tes Higher-Order Thinking Skill pada Pembelajaran Tematik Berbasis Outdoor Learning di Sekolah Dasar Kelas IV. Indonesian Journal of Primary Education, 2(1), 87-96.

Hendryadi, H. (2017). Validitas isi: tahap awal pengembangan kuesioner. Jurnal Riset Manajemen dan Bisnis (JRMB) Fakultas Ekonomi UNIAT, 2(2), 169-178.

Hikmah, H., & Amin, N. (2019). Pengembangan Instrumen untuk Mengukur Kemampuan Berpikir Tingkat Tinggi dalam Mata Pelajaran Matematika di SMA Kabupaten Majene. Saintifik: Jurnal Matematika, Sains, dan Pembelajarannya, 5(1), 1-7.

Imanuddin, T. N. F. (2015). Analisis Tingkat Kognitif Soal Apersepsi Pada Buku Siswa Matematika SMP/MTs Kelas VII Kurikulum 2013 Berdasarkan Taksonomi Bloom.

Mardapi Djemari.2016. Pengukuran, Penilaian, dan Evaluasi Pendidikan. Yogjakarta: Pustaka Pelajar.

Munadi, S. (2010, June). Analisis Kualitas Soal Untuk Penilaian Aspek Afektif. Makalah Disampaikan pada acara Workshop Penyuisunan Instrumen Evaluasi Afeektif mata kuliah Pengembangan Kepribadian, diselenggarakan pada tanggal.

Nugroho, B. S., Djuniadi, D., & Rusilowati, A. (2016). Pengembangan Penilaian Kinerja Menggambar Teknik Potongan di SMK pada Kurikulum 2013. Journal of Research and

Educational Research Evaluation, 5(1), 01-07.

Pulungan, D. A. (2014). Pengembangan Instrumen Tes Literasi Matematika Model Pisa. Journal of Research and Educational Research Evaluation, 3(2).

Retnawati Heri.2015.Validitas, Relibiltas, & Karakteristik Butir "Panduan untuk peneliti, Mahasiswa, dan Psikometrian". Yogyakarta: Parama Publishing.

Sa'idah, N., Yulistianti, H. D., & Megawati, E. (2019). Analisis Instrumen Tes Higher Order Thinking Matematika SMP. *Jurnal Pendidikan Matematika*, *13*(1), 41-54.

Sugiyono, M. (2015). *Penelitian & pengembangan (Research and Development/R&D).* Bandung: Penerbit Alfabeta.

Suhaesti Julianingsih, S. J., Undang Rosidin, U. R., & Ismu Wahyudi, I. W. (2017). Pengembangan instrumen asesmen hots untuk mengukur dimensi pengetahuan IPA siswa di SMP. *Jurnal Pembelajaran Fisika*, *5*(3).

Sujarwanto, S., & Rusilowati, A. (2015). Pengembangan Instrumen Performance Assessment Berpendekatan Scientific Pada Tema Kalor Dan Perpindahannya. *Unnes Science Education Journal*, *4*(1).

Wanda, V. N., Yusmin, E., & Nursangaji, A.(2019) Pengembangan Instrumen Tes Hots Berdasarkan Taksonomi Bloom Dalam Materi Trigonometri. *Jurnal Pendidikan dan Pembelajaran Khatulistiwa*, *8*(9).