Research Article

Validity and Reliability Analysis of HOTS Multiple Choice Questions in a Chemistry Course at a Senior High School

Anis Syafitri*, Murniaty Simorangkir, Ajat Sudrajat

Chemistry Education, Universitas Negeri Medan, Jl. Willem Iskandar Pasar V Medan, North Sumatra 20221, Indonesia

ORCID
Murniaty Simorangkir: https://orcid.org/0000-0002-3009-4090

Corresponding Author: Anis Syafitri; email: syafitrianis@gmail.com

Published: 3 April 2024

Publishing services provided by Knowledge E

Selection and Peer-review under the responsibility of the ICMScE Conference Committee.

Abstract.
This study examined the validity and reliability of a newly developed multiple-choice evaluation system that measured students' higher-order thinking skills (HOTS). The instrument test consisted of 45 multiple-choice items and was developed based on the cognitive domain of Bloom's Taxonomy. A quantitative method was used. It consisted of three phases: Content Validity by inter-rater agreement, Construct Validity by principal component analysis (PCA), and Reliability shown by Chronbach's alpha. The content validity by inter-rater agreement found that the instrument was categorized as valid. The construct validity by PCA found that each item in the evaluation instrument measured one-dimensionality, which is good to be used as an evaluation instrument test. The reliability was established to be a high degree with Chronbach's Alpha being 0.94. From the result of this study, a valid and reliable HOTS multiple-choice item evaluation instrument has been produced and is ready to be tested in a small sample to examine its empirical quality.

Keywords: validity, reliability, multiple-choice, evaluation system

1. INTRODUCTION

There are three aspects which assess in students' evaluation process such as aspect of knowledge, attitudes, and skills [1]. Knowledge is dominantly used as an aspect to determine completeness in the learning process during the semester at school. The policy of the Ministry of Education and Culture in 2018, in order to improve the quality of education, learning must be integrated with character education and based on Higher Order Thinking Skills (HOTS) [2]. According to Bloom's revised taxonomy, thinking skills are divided into two forms, namely Low Order Thinking Skills (LOTS), which are remembering, understanding, and applying. Then high order thinking skills (HOTS), such as analyzing, evaluating, and creating [3]. In order to enhance the ability

How to cite this article: Anis Syafitri*, Murniaty Simorangkir, Ajat Sudrajat, (2024), "Validity and Reliability Analysis of HOTS Multiple Choice Questions in a Chemistry Course at a Senior High School" in *International Conference On Mathematics And Science Education*, KnE Social Sciences, pages 367–376. DOI 10.18502/kss.v9i8.15568

of thinking, students are required to practice by doing HOTS questions, thus stimulating students to think more complex and advance [4, 5].

HOTS application in learning and assessment can improve students' way of thinking and skills [6]. The application of the HOTS assessment on the knowledge aspect is about to influence students' critical thinking skills [7]. Students need to improve HOTS especially in the ability to analyze and create in order to increase students' creativity in science [8]. However, according to the fact at schools, students' achievement and thinking ability is sill low. Based on the result in PISA 2018, Indonesia is ranked 74th out of 79 countries. This case indicates generally, the students' thinking skills in Indonesia are in a low category [9]. The main factor which caused it, is because the lack of HOTS instrument[10]. In accordance with this finding, students rarely practice by using HOTS questions. The questions used are limited to measuring low-level ability.

Therefore, 45 HOTS multiple-choice items have been developed which are used to measure the students' level of thinking ability. Multiple choice is an objective test which consists of five choices with one correct answer. The certainty of an instrument can be assessed from its validity and reliability [11]. Validity informs about the accuracy of the data collected from the certain field so that it accurately measures what should be measured [12]. In contrast reliability interested in the stability of the results if it is done with repetition and must measure the same construct [13].

The purpose of this study is to investigate the content validity using inter-rater agreement, construct validity using principal component analysis (PCA), and reliability using Cronbach alpha, coefficient. The results of this analysis will produce information about the quality of the HOTS question before being used as the instrument in the actual class.

2. RESEARCH METHOD

This study aimed to establish a valid and reliable HOTS multiple-choice questions. To determine the quality of the item, the prerequisite of validity and reliability must accomplish [14]. Hence, this study used the modified method [15] which consist of three phases: Content Validity, Construct Validity, and Reliability.

2.1. Participants

Participants were consisting of 151 senior high school students grade 12 in Langkat. Students were selected using the Simple Random Sampling (SRS) technique, which is done by determining a sample that meets certain criteria from the population [16]. The criteria used to privileged students were student in grade 12 with specialization in science from 2 high schools in Langkat who completed chemistry material in grade 11.

2.2. Procedure

Validity and reliability of the multiple-choice evaluation instrument were analyzed through 3 phases: Phase 1 – content validity by analyzing the results of the inter-rater agreement, Phase 2 – construct validity using principal component analysis (PCA), and Phase 3 – reliability was analyzed using Cronbach's alpha. Data analysis was executed using computer programs: Statistical Program for Social Science (SPSS) version 25.0 and Winsteps.

2.3. Phase 1: Content Validity

In this phase, the instrument was tested for content validity by 10 chemistry experts consists of 2 lecturers and 8 teachers. The experts are chosen based on the criteria who have been teaching for more than 5 years and considered to understand the arrangement of evaluation instruments. Data collection in content validity is taken using a Likert questionnaire. The questionnaire ranges from 1 to 4 with the interpretation 1 very lack, 2 lack, 3 good, 4 very good. The content validity was analyzed using descriptive quantitative analysis. The average results are interpreted as the expert agreement in a certain range of categories. Quantitative descriptive research is a portray of the research problem through a description of a situation or the need for an explanation of the relationship between variables [16].

2.4. Phase 2: Construct Validity

The analysis of construct validity used PCA. PCA was accomplished to ensure that the questions measure one dimension (unidimensionality), the dimension of knowledge. The students' answers were used as data in PCA analysis. There are some pre-requirements in order to determine PCA, that are meet the value of Kaiser-Mayer-Olkin Measure of Sampling Adequacy (KMO-MSA), Barlet's Test of Sphericity, and has a strong correlation proven by Anti Image Correlation (AIC). These tests are done by using SPSS program.

2.5. Phase 3: Reliability

The reliability test was examined from the Cronbach alpha value using the Winsteps program. This test identified the interaction between person and item as a whole [17]. To find out the level of respondent consistency, it can be seen from the person reliability results. Meanwhile, to see the quality of each item in the instrument was determined from the item reliability value.

3. RESULT AND DISCUSSION

There are 3 steps that have been done in order to determine validity and reliability of HOTS multiple-choice questions: (1) Content validity; (2) Construct validity; and (3) Reliability.

3.1. Phase 1: Content Validity

The HOTS evaluation instrument consists of 45 questions and had tested on the experts, that is lecturers and teachers. The data is collected by a questionnaire that assessed some aspects that contain material, construction, and language. The content validity was obtained from the average of expert agreement based on the questionnaire result. The finding in the inter-rater agreement is the same as the final average value. This final average shows the agreement level of the experts towards the instrument. Content validity recapitulation showed in Table 1.

Table 1: Average score of content validity.

Measured Aspects	Lecturers	Teachers	Total
Material	3.65	3.56	3.61
Construction	3.65	3.63	3.64
Language	3.69	3.50	3.60
Total Average	3.66	3.56	3.61

Table 1 shows the validity result by the lecturers ranges from 3.65 to 3.69. On the other hand, the validity result by teachers is around 3.50 to 3.65. The total average of inter-rater agreement is worth 3.61 which is categorized as valid and needs no revision. This result shows that the HOTS multiple-choice questions are decent as an instrument used in the actual class. The content validity is done to find out the worthiness of the instrument before it used in actual area [18].

3.2. Phase 2: Construct Validity

Construct validity was analyzed using PCA. PCA's finding describes how many dimensions which measured by a measurement tool. There are pre-requirements that have to be performed before determining PCA. Some pre-requirements categorized as success if the KMO-MSA > 0.5 and Bartlet's Test of Sphericity < 0.05 [19]. Then there is AIC, which can be fulfilled if the value > 0.5 per item. The result of KMO-MSA and Bartlet's Test of Sphericity is as shown in Table 2.

Table 2: KMO–MSA and Bartlet's test.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy. (KMO--MSA)	Bartlet's Test of Sphericity (Sig.)
0.905	0.000

Table 2 shown KMO–MSA and Bartlet's Test of Sphericity have accomplished. The result of KMO is higher than the standard and Bartlet's Test of Sphericity is lower than the standard. Lastly, the executing of AIC using SPSS. AIC is a partial correlation value between two variables with regard to the other variable is constant. The summary of AIC for each item is in Table 3.

Table 3: Anti-image correlation (aic) for each item.

Item	AIC	Item	AIC	Item	AIC	Item	AIC	Item	AIC
Q1	.865^a	Q11	.625^a	Q21	.921^a	Q31	.927^a	Q31	.948^a
Q2	.896^a	Q12	.944^a	Q22	.888^a	Q32	.943^a	Q32	.943^a
Q3	.926^a	Q13	.926^a	Q23	.885^a	Q33	.898^a	Q33	.853^a
Q4	.911^a	Q14	.895^a	Q24	.545^a	Q34	.934^a	Q34	.858^a
Q5	.876^a	Q15	.928^a	Q25	.920^a	Q35	.953^a	Q35	.926^a
Q6	.929^a	Q16	.842^a	Q26	.912^a	Q36	.921^a		
Q7	.839^a	Q17	.501^a	Q27	.916^a	Q37	.930^a		
Q8	.878^a	Q18	.922^a	Q28	.822^a	Q38	.775^a		
Q9	.930^a	Q19	.836^a	Q29	.917^a	Q39	.900^a		
Q10	.864^a	Q20	.920^a	Q30	.866^a	Q40	.945^a		

Table 3 shows the result of AIC which is marked with "a". All AIC values in Table 3 > 0.5. Therefore, the AIC requirement is fulfilled. If there is an item that does not meet the AIC criteria, that item should be removed and the pre-requirement test should be repeated. But, if there is not, validity construct determination can proceed to PCA [20].

PCA determination is done to analyze the dominant factors measured by the HOTS multiple-choice questions. The analysis was conducted from the result of initial eigenvalue shown in Total Variance Explained in the SPSS output. If the total the initial eigenvalue < 1, it interprets that the factor can not explain the variable well. The percentage of factors that can explain the variance of 45 items is around 64.74% and collected as 10 components. The details of total variance which is explained by the 10 components from the eigenvalue are shown in Table 4.

Table 4: Percentages of varians which explained by 10 factors from initial eigenvalue.

Comp.	Total	%Varians	%Cummulatives	Comp.	Total	%Varians	%Cummulatives
1	14.815	32.922	32.922	6	1.423	3.161	54.245
2	2.506	5.569	38.490	7	1.278	2.840	57.086
3	2.350	5.223	43.713	8	1.219	2.709	59.795
4	1.819	4.042	47.755	9	1.163	2.585	62.379
5	1.498	3.329	51.084	10	1.073	2.385	64.764

Table 4 shows from PCA, there are 10 components contained in the instrument with a percentage of around 64.764% that can explain 45 items' variance. According to Table 4, it is known that the first component can explain 32.922% of the total components. This indicates there is one dominant factor in the HOTS evaluation instrument because it has the most dominant variance percentage. Along with this finding, it is known that there is a dominant factor that basing students' responses in order to complete the instrument. Domination of one factor can affect students to act [21]. In this study, the one action that is influenced by the dominant factor is the high order thinking skill of the students.

The comparison of eigenvalue and sum of all components is described by scree plot graphic in Figure 1. This analysis can confirm the unidimensionality test is seen from the scree plot in Figure 1. In Figure 1 extreme steepness is formed between eigenvalues 1 and 2, while the other eigenvalues only form small steepness. The sharp decline in Figure 1 described the factor that dominated the variances. Figure 1 can be used as the indicator of the dominant factor in the case that there is a rapid difference between components 1 and 2 [22].

Figure 1 shows that there is an extremely steep between the first and the second eigenvalue, on the other hand, the other eigenvalue seems to have a slight change. This significant step in the scree plot graphic is interpreted as a measurement of one dominant factor. The PCA result revealed that there is a fit between the model and the empirical data [23].

3.3. Reliability

Determination of reliability describes the consistency of the instrument in the measurement if it is used repeatedly. The overall reliability value can be seen from Cronbach

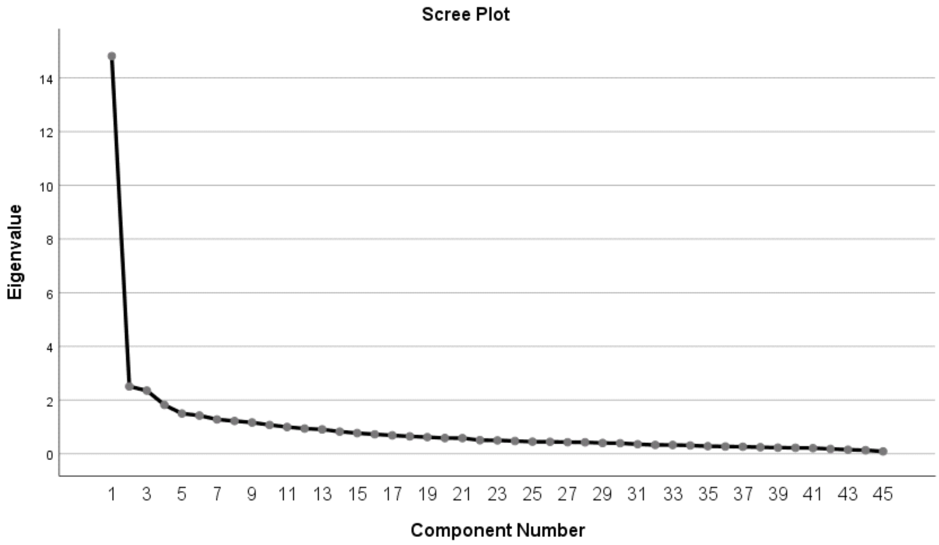

Figure 1: Scree plot: ratio between the number of components and eigenvalue.

alpha. The determination of the consistency level of respondents can be seen from the results of Person Reliability. Meanwhile, to identify the quality for each item in the instrument, it was determined from the Item Reliability. The results of the analysis showed in Table 5.

Table 5: Reliability result.

Reliability	Result	Category
Chronbach's Alpha	0.76	Good
Person Reliability	0.68	Enough
Item Reliability	0.90	Very Good

Table 5 shows that the instrument is reliable with the criterion of Chronbach alpha is good which means the interaction between students and items is acceptable. Person reliability is enough, and the item reliability is very good. The consistency of HOTS multiple-choice questions was established by comparing the outcome and students' skills.

Based on the results of the analysis in phases 1 to 3, 45 questions are categorized as good, from the results of content validity, construct validity, and reliability. These HOTS multiple-choice questions are ready to be tested on a limited sample to examine its empirical quality. So as the product of this study, these HOTS questions will be ready to be used in the real class.

4. CONCLUSION

The examination of validity and reliability of a HOTS multiple-choice questions had accomplished. It took three phases such as content validity by inter-rater agreement, construct validity by principal component analysis (PCA), and reliability shown by Chronbach alpha. Firstly, the content validity of the instrument was examined by inter-rater agreement (expert judgment) and result as valid. Secondly, a good evaluation instrument has to construct well and it results each item measured one-dimensionality which good to be used as an evaluation instrument test. Lastly, the interaction between the students' and items is acceptable shown by the chronbach's alpha with 0.76 and 45 HOTS multiple-choice questions are ready to be tested on a limited sample to examine its empirical quality. From these findings, a valid and reliable HOTS multiple-choice questions has been produced.

References

[1] Syahida A, Irwandi D. Analisis keterampilan berpikir tingkat tinggi pada soal ujian nasional kimia. Edusains. 2015;7(1):77–87.

[2] Wiwik S. Buku Penilaian Berorientasi Higher Order Thinking Skills. 2015.

[3] Bloom BS, Krathwohl DR, Taxonomy of educational objectives: The classification of educational goals. Book 1, Cognitive domain. Logman; 2020.

[4] Nurwahidah I. Pengembangan soal penalaran model timss untuk mengukur high order thinking (HOT). Thabiea : Journal of Natural Science Teaching. 2018;1(1):20.

[5] Subia GS, Marcos MC, Pascual LE, Tomas AV, Liangco MM. Cognitive levels as measure of higher-order thinking skills in senior high school mathematics of science, technology, engineering and mathematics (STEM) graduates. Technology Reports of Kansai University. 2020;62(3):261–8.

[6] Brookhart SM. How to assess higher-order thinking skills in your classroom. Ascd; 2010.

[7] Nurhayati S, Ningrum RT. Influence of cognitive assessment instrument based higher order thinking skill toward students critical thinking skill. Proceeding of ICMSE. 2016;3(1).

[8] Saido GM, Siraj S, Nordin AB, Al Amedy OS. Higher order thinking skills among secondary school students in science learning. The Malaysian Online Journal of Educational Science. 2015;3(3):13–20.

[9] Hewi L, Shaleh M. Refleksi hasil pisa (the programme for international student assesment): upaya perbaikan bertumpu pada pendidikan anak usia dini). Jurnal Golden Age. 2020;4(01):30–41.

[10] Ghani IB, Ibrahim NH, Yahaya NA, Surif J. Enhancing students' HOTS in laboratory educational activity by using concept map as an alternative assessment tool. Chem Educ Res Pract. 2017;18(4):849–74.

[11] Heale R, Twycross A. Validity and reliability in quantitative studies. Evid Based Nurs. 2015 Jul;18(3):66–7.

[12] Hulteen RM, Barnett LM, True L, Lander NJ, Del Pozo Cruz B, Lonsdale C. Validity and reliability evidence for motor competence assessments in children and adolescents: A systematic review. J Sports Sci. 2020 Aug;38(15):1717–98.

[13] Taherdoost H. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in research. How to test the validation of a questionnaire/survey in research. 2016. https://doi.org/10.2139/ssrn.3205040.

[14] Sudijono A. "Pengantar evaluasi pendidikan.," p. 2001.

[15] Barak M, Watted A, Haick H. Establishing the validity and reliability of a modified tool for assessing innovative thinking of engineering students. Assess Eval High Educ. 2020;45(2):212–23.

[16] Creswell JW. Educational research: planning, conducting, and evaluating quantitative and qualitative research. 2012.

[17] Eka S, Purba D. Analisis model Rasch instrumen tes prestasi pada mata pelajaran dasar dan pengukuran listrik A Rasch model analysis of instrument achievement test on basic electrical lesson and electrical measurements. Jurnal Penelitian dan Evaluasi Pendidikan. 2018;6(2):142–147.

[18] Bus Umar H. Principal component analysis (pca) dan aplikasinya dengan spss. Jurnal Kesehatan Masyarakat Andalas. 2009;3(2):97–101.

[19] Alfarisa F, Purnama DN. Analisis butir soal ulangan akhir semester mata pelajaran ekonomi sma menggunakan rasch model. 2019;11(2).

[20] Ridho A. Karakteristik psikometrik tes berdasarkan pendekatan teori tes klasik dan teori respon aitem. Insan Media Psikologi. 2007;9(2):83–104.

[21] Santoso A, Kartianom K, Kassymova GK. Kualitas butir bank soal statistika (Studi kasus: instrumen ujian akhir mata kuliah statistika Universitas Terbuka). Jurnal Riset Pendidikan Matematika. 2019;6(2):165–76.

[22] Morad S, Ragonis N, Barak M. The validity and reliability of a tool for measuring educational innovative thinking competencies. Teach Teach Educ. 2021; 97:103193.

[23] Solihatun S, Rangka IB, Ratnasari D. Measuring of student learning performance based on geometry test for middle class in elementary school using dichotomous Rasch analysis. Journal of Physics: Conference Series. 2019;1157(3). https://doi.org/10.1088/1742-6596/1157/3/032086.