

The Development of Two-Tier Multiple Choice Instruments to Measure Higher Order Thinking Skills Bloomian

Pangestika Nur Afnia¹, Edi Istiyono^{2*}

¹Program Pascasarjana Universitas Negeri Yogyakarta

²Jurusan Pendidikan Fisika Universitas Negeri Yogyakarta

¹ tikaafnia@gmail.com, ^{2*} edi_istiyono@uny.ac.id

Abstract: One of the educational problems is the quality of assessment and evaluation instruments. This study aims to develop two-tier multiple choice instruments to measure Higher Order Thinking Skills (HOTS) Bloomian in acid and base material. Bloomian's high order thinking skills consist of three aspects, they are analyzing, evaluating, and creating. The instrument consisted of two sets of tests, each containing 35 items including seven anchor questions that had been validated by two chemistry education experts and a measurement expert. The instrument was tested on 172 students spread across five high schools in the city of Yogyakarta. The results of the trials were analyzed and showed that as many as 63 questions fit with infit MNSQ 0.77 – 1.29. The level of difficulty in the test instrument ranges from -0.85 and 0.76. The instrument reliability is 0.71 which means it is included in the good category and qualified the requirements to measure high order thinking skills of high school students.

Keywords: *test instruments, higher order thinking skills, two-tier multiple choices*

INTRODUCTION

One of the problems in education is the low quality of assessment and evaluation instruments. Assessment is the process of gathering information about student learning processes (Moria, Refnaldi, & Zaim, 2017). Information from the assessment process is needed for decision making regarding students' learning abilities and achievements (Kankam, Bordoh, Eshun, Bassaw, & Korang, 2014). Assessment helps teachers classify students into specific groups, improve teaching methods, and measure students' readiness both attitude, mental, and material readiness (Retnawati, Hadi, & Nugraha, 2016). The inability of teachers to conduct effective assessments will bring difficulties for teachers in maximizing learning in the classroom (Nenty, Adedoyin, Odili, & Major, 2007). The success of assessment process depends on the selection and use of appropriate and effective procedures, as well as on the correct interpretation of student achievement (Bordoh, Eshun, Quarshie, Bassaw, & Kwarteng, 2015).

Teachers in schools make test questions not to measure students' thinking skills but merely measure the achievement of learning objectives. Thinking skills are divided into lower order thinking skills (LOTS) and higher order thinking skills (HOTS). Indonesian educators are more familiar with the term Bloom's taxonomy in the preparation of tests and curricula. Dimensions of cognitive processes in Bloomian's taxonomy consist of recalling, understanding, applying, analyzing, evaluating, and creating. The ability to analyze, evaluate and create is classified as a higher order thinking skills. The ever-changing world requires students to develop higher-order thinking skills, such as critical system thinking, decision making, and problem solving (Miri, David, & Uri, 2007).

The higher order thinking skills for structured inquiry are best obtained if there is an evaluation system that measures high-level complex skills rather than just remembering simple facts (Hopson et al.: 2014). The results of the Program for International Student Assessment

(PISA), Indonesia was in 62 rank which is lower than other Southeast Asian countries such as Singapore and Vietnam. PISA is a program organized by the Organization for Economic Cooperation and Development (OECD) to evaluate literacy, scientific and mathematical abilities which aims to determine the ability of 15-year-old students ability and skill. The PISA questions are classified as higher order thinking skills (HOTS) in the Bloomian taxonomy (Setiawan, Dafik, & Lestari, 2014). This shows that the ability to analyze, evaluate, and the creation of students in Indonesia is classified as very low.

One of the factors causing the Indonesia's achievement in PISA is the low ability of teachers to develop assessment instruments for students' higher order thinking skills. Implementation of higher order thinking skills (HOTS) in the assessment aspect aims to assess the ability of students in solving high-level questions (Abdullah, Mokhtar, Halim, Ali, Tahir, & Kohar, 2017). Good assessment will encourage educators to determine quality learning strategies and motivate students to learn better. HOTS questions are items that assess cognitive skills for analyzing, evaluating, and creating (Mohamed & Lebar, 2017).

HOTS assessment questions can be in the form of multiple choice, structured questions, essays and performance assessment (Brookhart, 2010). The multiple choice test instrument is a form of test that is often used because it is practical and easy to assess (Erifianti, Istiyono, & Kuswanto, 2019). Multiple choice test students who answer incorrectly are considered not knowing the answer to the question and students who answer correctly are considered to know the answer to the question even though by guessing the answer (Karandikar, 2010). Therefore, the possibility of students guessing the correct answer is one of the weaknesses of multiple choice questions.

Two-tier multiple choice questions are another form of multiple choice question instruments where there are two answers in one item. The first answer is the answer to the problem, while the second answer is the reason that underlies students in answering the first answer. Therefore, to answer correctly, students' understanding and thinking ability are needed. Therefore, Two-tier multiple choice questions as assessment instruments are developed to measure higher order thinking skills os students.

METHODS

This is a development research using Thiagarajan 4D development model which consisting of 4 stages; define, design, develop, and disseminate. Define the stage is a defining phase that aims to collect information related to the development research. At the design stage the test instrument includes the determination of the test objectives, the development of the form of the test instrument, determination of the test material, preparation of the test blue print, and guidelines of test instruments was written. The development stage consists of two activities. The first is validation process to assess the product feasibility and the second is empirical testing to students. The dissemination stage is the final stage of product development such that the product can used by teacher.

The development of test instruments in this research is limited to chemistry subjects in acid-base material, hydrolysis, and buffer solutions. The instrument contained 63 items with seven anchor questions divided into two instruments set. The existence of anchor items is intended for equating the test results so that the test results can be compared (Istiyono, et al: 2014).

RESULTS AND DISCUSSION

Higher order thinking skills (HOTS) test instrument Bloomian revised Anderson and Krathwohl developed contains three aspects, namely analyzing (C4), evaluating (C5), and creating (C6). The test instrument grid is arranged based on these three aspects which are then derived into 33 indicators. The indicators that have been compiled are then developed into two instruments set, which each of them contains 35 item.

Table 1. HOTS Indicator Matrix

HOTS Indicators	HOTS Subaspect	Topics
Analyze	Diffentiating	Acid-Base Hydrolisis Buffer Solution
	Organizing	
	Attributing	
Evaluate	Checking	
	Critiquing	
Create	Generating	
	Planning	
	Producing	

Two experts in chemistry education and one measurement expert were involved in the preparation of the test instrument. This experts validation is used to obtain the content validity. Validation results are shows as follows:

Table 2. Aiken Result Validity

Item Number	V Aiken	validity
7A, 24A, 33, 35, 5B, 8B, 17B, 19B, 23B	0,67	Valid
1A, 2A, 4A, 6A, 8A, 9A, 12A, 13A, 14A, 17A, 22A, 26, 28, 29A, 1B, 4B, 7B, 13B, 14B, 24B, 25B, 29B	0,78	Valid
3A, 11A, 15A, 16A, 18A, 19A, 20, 21, 23A, 25A, 27, 30A, 31A, 32A, 2B, 3B, 6B, 9B, 10, 11B, 12B, 16B, 18B, 22B, 30B, 31B, 32B, 34B	0,89	Valid
5A, 10A, 15B,	1,00	Valid

The revised test instrument was then tested on 172 students in five Yogyakarta high schools. The results were analyzed using the item response theory and they're must fit with item response theory assumptions like unidimensional, local independence, and parametric invariance. Unidimensional assumptions must be fulfilled to prove that each item only measures one ability. The assumption test by using factors analysis that produce Kaiser-Meyer-Olkin (KMO) output and Eigenvalue that shown in Figure 1 and Figure 2. The KMO results is 0.982 which means that the sample size used was fit with the assumptions. The dimensions of test instrument can also be seen through the eigenvalues in the scree plot in figure 3 which shows one dominant component.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	.982
Approx. Chi-Square	31017.124
Bartlett's Test of Sphericity	df
	1953
Sig.	.000

Figure 1. KMO and Bartlett's Test

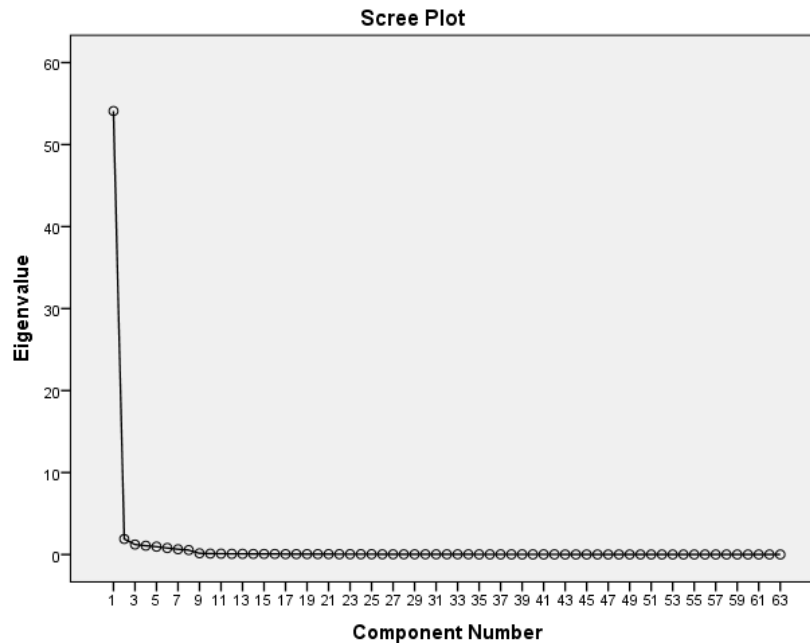


Figure 2. Scree Plot

If unidimensional assumption was fulfilled, the local independence was also but its not applicable with parametric invariance. The parametric invariance proved that the item not depend on the students ability parameters distribution and the ability parameters distribution not depend on the item test. The parametric invariance shows in Figure 3 and Figure 4. The results of the scree plots show that the dots form linear line which concluded the instrument was fulfilled the invariance assumption.

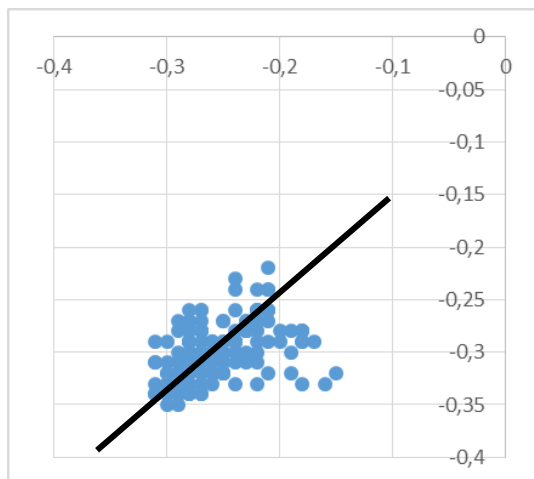


Figure 3. Scree Plot of Ability Invariance

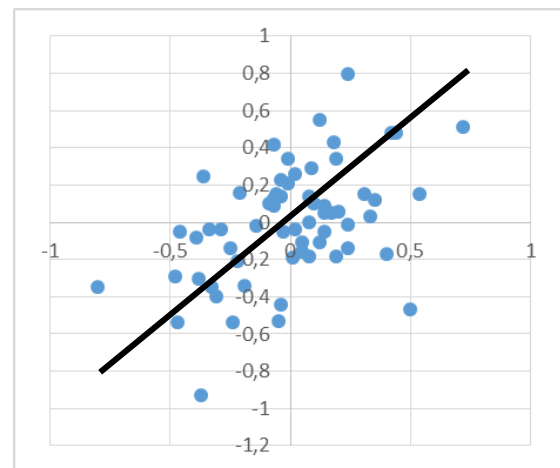


Figure 4. Scree Plot of Difficulty Invariance

Analysis to determine the difficulty level of the items and instrument reliability was carried out using software Quest. The Bloomian high-level skills test (HOTS) instrument in this study was developed in the form of two-tier multiple choice item. The criteria for scoring a higher order thinking skill (HOTS) test in Bloomian are made with four scoring categories. Category 1 is students answer wrong questions and wrong argument will get 0 score. Category 2 is students

answer the right questions and wrong argument will get 1 score. Category 3 is students answer the wrong questions and right argument will get 2 score. While category 4 is students answer the right questions and the right argument will get 3 score. Therefore the level of thinking skills of students can be known from these categories.

The characteristics of the items in the item response theory consist of discriminant, difficulty and guessing correct answer (Sudaryono: 2011). In this development study, researchers used an item response theory (IRT) with the Rasch model approach which only have item difficulty. Difficulty is defined as the probability of answering a question correctly at a certain level of ability. The easier the item is, the greater the probability that the test taker will answer correctly on the item. The function of the level of difficulty of the item relates to the purpose held of the test (Aiken: 1994).

Table 3. Difficulty Results of items

ITEM Code	DIFF	ITEM Code	DIFF	ITEM Code	DIFF	ITEM Code	DIFF	ITEM Code	DIFF	ITEM Code	DIFF	ITEM Code	DIFF
A1	-0,51	A10	0,13	A19	0,31	28	0,14	B2	-0,44	B11	0,19	B22	-0,04
A2	-0,32	A11	0,47	20	0,03	A29	0,49	B3	-0,27	B12	0,49	B23	-0,21
A3	-0,25	A12	0,17	21	-0,09	A30	0,45	B4	0,12	B13	0,17	B24	-0,25
A4	-0,35	A13	0,14	A22	0,36	A31	0,51	B5	-0,22	B14	0,08	B25	-0,07
A5	-0,85	A14	-0,08	A23	-0,11	A32	0,16	B6	-0,46	B15	-0,21	B29	0,27
A6	-0,33	A15	-0,04	A24	0,08	33	0,14	B7	-0,04	B16	-0,1	B30	-0,09
A7	-0,37	A16	0,2	A25	0,17	A34	0,76	B8	-0,45	B17	0,07	B31	-0,1
A8	-0,07	A17	-0,07	26	-0,05	35	0,5	B9	-0,04	B18	0,08	B32	0,04
A9	-0,43	A18	0,16	27	0,14	B1	-0,47	B10	0,07	B19	0,07	B34	0,18

INFIT	.56	.63	.71	.83	1.00	1.20	1.40	1.60	1.80
MNSQ									
1 item 1					*				
2 item 2					*				
3 item 3					*				
4 item 4					*				
5 item 5					*				
6 item 6					*				
7 item 7					*				
8 item 8					*				
9 item 9					*				
10 item 10					*				
11 item 11					*				
12 item 12					*				
13 item 13					*				
14 item 14					*				
15 item 15					*				
16 item 16					*				
17 item 17					*				
18 item 18					*				
19 item 19					*				
20 item 20					*				
21 item 21					*				
22 item 22					*				
23 item 23					*				
24 item 24					*				
25 item 25					*				
26 item 26					*				
27 item 27					*				
28 item 28					*				
29 item 29					*				

30 item 30	.	*	.
31 item 31	.	*	.
32 item 32	.	*	.
33 item 33	.	*	.
34 item 34	.	*	.
35 item 35	.	*	.
36 item 36	.	*	.
37 item 37	.	*	.
38 item 38	.	*	.
39 item 39	.	*	.
40 item 40	.	*	.
41 item 41	.	*	.
42 item 42	.	*	.
43 item 43	.	*	.
44 item 44	.	*	.
45 item 45	.	*	.
46 item 46	.	*	.
47 item 47	.	*	.
48 item 48	.	*	.
49 item 49	.	*	.
50 item 50	.	*	.
51 item 51	.	*	.
52 item 52	.	*	.
53 item 53	.	*	.
54 item 54	.	*	.
55 item 55	.	*	.
56 item 56	.	*	.
57 item 57	.	*	.
58 item 58	.	*	.
59 item 59	.	*	.
60 item 60	.	*	.
61 item 61	.	*	.
62 item 62	.	*	.
63 item 63	.	*	.

Figure 5. Instrument INFIT Meansquare

Summary of item Estimates

=====

Mean	-.26
SD	.27
SD (adjusted)	.23
Reliability of estimate	.71

Figure 6. Instrument Reliability Estimates Output

The items in the Bloomian Higher Order Thinking Skills (HOTS) test instrument fit with INFIT MNSQ criteria at 0.77 to 1.29 so that it can be concluded that 63 fit items and can be used to measure higher order thinking skills empirically. Based on the analysis also obtained a reliability value of 0.71. This shows that the reliability level of the instrument is in the good category.

CONCLUSIONS

Based on the results of data analysis and discussion of the research that has been done, conclusions can be drawn as follows:

- 1) There are 63 items that fit with the criteria of INFIT MNSQ (0.77 to 1.29).
- 2) The average level of difficulty of the instruments ranging from -0,85 to 0,76

3) The instrument reliability coefficient is 0.71 that means in good category.

REFERENCES

- Abdullah, A.H., Mokhtar, M., Halim, N.D.A., Ali D.F., Tahir, L.M., & Kohar, U.H.A. (2017). Mathematics teachers' level of knowledge and practice on the implementation of higher-order thinking skills (HOTS). *EURASIA Journal of Mathematics Science and Technology Education*, No. 13(1), Hal. 3-17
- Aiken, L.R. (1980). Three coefficient of analyzing the reliability and validity of ratings. *Internal of Educational and Psychological Measurement*, 40, 955-967
- Bordoh, A., Eshun, I., Quarshie, A.M., Bassaw, T.K., & Kwarteng, P. (2015). Social Studies Teachers' Knowledge Base in Authentic Assessment in Selected Senior High Schools in the Central Region of Ghana. *Journal of Social Sciences and Humanities*, Vol. 1, No. 3, Hal. 249-257
- Brookhart, S.M. (2010). *How to assess higher-order thinking skills in your classroom*. Alexandria, Virginia: ASCD
- Erfianti, L., Istiyono, E. & Kuswanto, H. (2019). Developing lup instrument test to measure higher order thinking skills (HOTS) bloomian for senior high school students. *International Journal of Educational Research Review*, Vol. 4 (3), Hal. 320-329
- Hopson, M.H., Simms, R.L., & Knezek, G.A. (2014). Using a Technology-Enriched Environment to Improve Higher-Order Thinking Skills. *Journal of Research on Technology in Education*, 34 (2), 109-119
- Istiyono, Edi. (2013). Tes Kemampuan Berpikir Tingkat Tinggi Fisika di SMA Langkah Pengembangan Dan Karakteristiknya. *Artikel Penelitian Disertasi Doktor*: Universitas Negeri Yogyakarta
- Istiyono, Edi., Djemari Mardapi., dan Suparno. (2014). Pengembangan Tes Kemampuan Berpikir Tingkat Tinggi Fisika (PysTHOTS) Peserta Didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan* Vol.18 (1), Hal. 1-12
- Kankam, B., Bordoh, A., Eshun, I., Bassaw, T.K., & Korang, F.Y. (2014). Teachers' perception of authentic assessment techniques practice in social studies lessons in senior high schools in Ghana. *International Journal of Educational Research and Information Science*, 1(4), Hal. 62-68
- Karandikar, R.L. (2010). On multiple choice tests and negative marking. *Current Science*, Vol. 99, No. 8, Hal. 1042 – 1045
- Miri, B., David, B.-C., & Uri, Z. (2007). Purposely teaching for the promotion of higher-order thinking skills: a case of critical thinking. *Research in science education*, 37 (4), Hal. 353-369
- Nenty, H.J., Adedoyin, O.O., Odili, J.N., & Major, T.E. (2007). Primary teacher's perceptions of classroom assessment practices as means of providing quality primary/basic education by Botswana and Nigeria. *Educational Research and Review*, Vol. 2 (4), Hal. 074-081
- OECD. (2015). *PISA 2015 Scientific Literacy Framework*. Paris: Organization for Economic Cooperation and Development.
- Retnawati, H., Hadi, S., & Nugraha, A.C. (2016). Vocational high school teachers' difficulties in implementing the assessment in curriculum 2013 in Yogyakarta Province of Indonesia. *International Journal of Instruction*, Vol. 9, No. 1, Hal. 33-48
- Setiawan, H., Dafik, Lestari, N. (2014). Soal matematika dalam pisa kaitannya dengan literasi

matematika dan keterampilan berpikir tingkat tinggi. *Prosiding Seminar Nasional Matematika, Universitas Jember, 19 November 2014, Hal. 241-244-251*