JIPI 5(2):172-185, 2021
Jurnal IPA dan Pembelajaran IPA
e-ISSN: 2620-553X p-ISSN: 2614-0500

Quantitative Analysis Towards Higher Order Thinking Skills of Chemistry Multiple Choice Questions for University Admission

Muhammad Reza[1*], Kana Puspita[2], Coryna Oktaviani[3]

[1]Program Studi Pendidikan Kimia, Fakultas Tarbiyah dan Keguruan, Universitas Islam Negeri Ar-Raniry, Banda Aceh, Indonesia

[2]Program Studi Pendidikan Kimia, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Syiah Kuala, Banda Aceh, Indonesia

[2]Program Studi Pendidikan Kimia, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Samudra, Langsa, Indonesia

*Email: muhammad.reza@ar-raniry.ac.id

DOI: 10.24815/jipi.v5i2.20508

Article History:	*Received: March 26, 2021*	*Revised: May 5, 2021*
	Accepted: May 25, 2021	*Published: June 16, 2021*

Abstract. Since curriculum 2013 required higher order thinking skill questions to be used in examination for university admission, the teachers must provide the kind of questions for learning evaluation at school. To achieve good quality, the questions must be analyzed qualitatively and quantitatively.This research aims to analyze the test of validity, reliability, correlation between difficulty index and discriminatory power, functioning and nonfunctioning distractors of chemistry multiple questions quantitatively which are frequently offered to students in national university admission. It used a descriptive-quantitative method and purposive sampling technique. Students of the acceleration class were selected purposely because they have been frequently answering the higher order thinking questions. Data were collected using 20 multiple choice questions developed from questions banks. Before questions used for students, the questions were assessed by two validators. The result of the validity test shows that 100% questions were valid with an average scale score of 0.93 (best valid). Further analysis used a reliability test to indicate the correlation of item score and total score of questions symbolized by Cronbach alpha of 0.94, 0.92 and 0.89 for aspect of content, construction, language or culture, respectively. The correlation between difficulty index and discriminatory power was reliable with correlation coefficient of 0.842. The final analysis is done towards distractor function, and it shows that overall 83.75% was functioning distractors. Then, nine questions or 45% met ideal expectations with four functioning distractors. Therefore, the most higher order thinking chemistry multiple questions have good quality or feasible to be used in university admission.

Keywords: higher order thinking, validity, reliability, difficulty index, discriminatory power and distractors

Introduction

Assessment is an important element in learning. Generally, it is carried out through tests, like questions and questionnaires. But, the most frequent test for learning is done by using questions. A common questions test is multiple-choice questions. Multiple choice questions are largely used as formal tests for evaluation or assessment in several fields of education (Begum, 2012; Douglas, et al., 2012; Butler, 2018). Also in Indonesia, this kind of question is used for national examinations either at school or for university admission.

Since 2013, the Indonesian government has been pushing teachers to improve the style of mutual questions developed for learning assessment. Due to the curriculum goal to achieve student higher order thinking skills, teachers are required to create or develop higher order thinking skills (HOTS) questions (Hamdi, et al., 2018). It helps students to master the examination with more complexity, like the test for university admission. One of the subjects that becomes test materials is chemistry. In addition, from 2018, the government rearranged the rules of examination for university admission by adding HOTS content inside the test (Harta, et al., 2020). Besides that, the national examination also contains higher order thinking skills because it is one of the important components of the 2013 Curriculum in Indonesia. In case, students' ability in HOTS is generally not achieved the expectations (Puspitasari & Nugroho, 2020). In addition, HOTS skill has strong correlation with critical thinking skill. The result of students' critical thinking skill can not be separated from learning designed by teachers (Sinurat at al., 2020; Isa et al., 2017; Rahmati et al., 2017; Suwono et al., 2017). So, to make students familiarize with the test of HOTS chemistry questions, teachers must develop HOTS questions for learning assessment at school. Therefore, students are able to enhance their analysis skills in order to solve the HOTS questions (Tsaparlis, 2020). By having HOTS skills, students will not only be able to analyze, but also to evaluate, and create innovation in solving problems (Ichsan, et al., 2019).

To obtain the better quality of HOTS questions, the development process must be completed by qualitative and quantitative analysis of the test item. The multiple-choice questions are analyzed using several methods, and one of the methods is item analysis to assess the quality of those items (KoÇdar, et al., 2016). The questions should contain several qualitative aspects such as concepts, construction and language. The proper construction of questions is a must in order to create effective multiple-choice questions. The effective questions have a good reliability and validity. Without reliability and validity test, the multiple-choice questions only measure the basic memory like to recall the factual knowledge rather than higher order thinking skills (Douglas, et al., 2012). When the questions stand with better validity and reliability, the multiple-choice questions become the most efficient evaluation form for assessment (Begum, 2012; DiBattista, et al., 2013; Dehnad, et al., 2014; Kingston, et al., 2012).

Besides qualitative analysis, quantitative analysis aspects that must be included for each of the test items are difficulty index, discriminatory power, distractor. Previous research done by Mahjabeen et al. (2017) showed 81% multiple-choice questions were average, while 2% and 17% were too easy and too difficult, respectively. Then, the similar result reported by Taib & Yusoff (2014) that the difficulty index of multiple-choice questions ranged from 0.67-0.79, which is represented as an average index (optimal level). The difficulty index is related to discriminatory power like reported by Pande et al. (2013) that discriminatory power has positive correlation with difficulty index, and the greatest discrimination was exihibited with easy or difficult question items. After that, the analysis of distractor function is important to complete discriminatory index in distinguishing high-performing and low-performing students. So, it is recommended to include distractor analysis for similar studies (KoÇdar, et al., 2016). Aside from difficulty, discriminatory power and distractors, every test item should archive other qualities like validity and reliability (Osundare and Omirin, 2016).

The analysis aims to examine the test item for minimum standard to be a good test instrumentation. If the questions still have the lack of both qualitative and quantitative aspects, the test item can be changed or removed. In addition, if the result during analysis is good, the questions are feasible to be used as test questions and can be stored for further examinations. Before the teachers start to develop the valid HOTS questions, they should take some questions reference for analysis. A question set of examination for university admission is mostly built by HOTS questions, so it can be used as reference and comparison materials. Therefore, researchers were interested to analyze the university admission questions test both quantitatively and qualitatively. So, it will ease teachers in developing HOTS questions with better quality based on analysis aspects.

Methods

This research used a descriptive-quantitative method. It was conducted in the acceleration class at SMAN Modal Bangsa, Aceh Province. Sample was selected using purposive sampling technique, because they frequently follow the test with higher order thinking skills. Data were collected using twenty multiple-choice questions developed from question banks. Before use, the questions were validated by experts in learning evaluation and assessment. Then, further analysis is done through a reliability test for aspects of content, construction, and language or culture. The samples are requested to answer the multiple choice questions, and their answer was analyzed quantitatively based on difficulty index, discriminatory power, functioning and nonfunctioning distractors. In addition, the correlation between difficulty index and discriminatory power was evaluated.

Difficulty index. Difficulty index is a benchmark for question quality used for learning evaluation. The index is good when the difficulty index is moderate. The values of difficulty index were calculated using equation (1) as follows:

$$P = \frac{\square}{\square\square} \qquad (1)$$

Where P is difficulty index, B is number of students who have the correct answer to the specific question number, and JS is the number of students (Sugiyono, 2009). To interpret the values of difficulty index, it follows the following Table 1 (Towns, 2004).

Table 1. Interpretation guideline for difficulty index

No	Difficulty index	Interpretation
1	Below 0.25	Difficult
2	0.25 – 0.75	Average
3	Above 0.75	Easy

Discriminatory power represents the ability of a question test to distinguish outstanding and non-outstanding students. The values of discriminatory power can be calculated using the equation (2):

$$D = P_A - P_B \qquad (2)$$

Where D is discriminatory power, P_A is the number of correct answers of high performing students, and P_B symbolizes the number of correct answers of low performing students (Sugiyono, 2009) After that, the values were classified into four interpretations as shown in Table 2 (Bhat & Prasad, 2021).

Table 2. Classification of discriminatory power

No	Discriminatory power values (D)	Classification
1	Above 0.35	Excellent
2	0.20 – 0.35	Good
3	Below 0.2	Poor

The Distractor function describes the way on how the options can be used as alternative answers by students. It is good to be functioning when an option is selected by 5% of the sample.

Validity. Validity test towards questions carried by two validators. The validation is done using a standard assessment with 18 question items divided into three aspects; content, construction, and language or culture. The questions are feasible if they meet eligibility criteria as shown in Table 3 (Hadisaputra, et al., 2020). Further test of validity used correlation coefficient between item score and total score.

Table 3. Validation criteria

No	Average assessment scale	Validity Level
1	0.80 - 1.00	Best Valid
2	0.66 - 0.79	Valid
3	0.56 - 0.65	Average
4	0.40 - 0.55	Less
5	0.30 - 0.39	Invalid

Reliability. To find out the reliability of the question, the questions were analyzed through three aspects, namely content, construction, and language or culture. The correlation between item score and total score was analyzed quantitatively using Cronbach alpha. The coefficient of correlation is generally accepted at values of 0.7 or 0.6 (Greithuijsen, et al., 2014)

Results and Discussion

Validity dan Reliability

All questions were analyzed qualitatively according to three aspects, such as concept, construction and language or culture by two validators. The content aspect focuses on relevant indicators and learning competencies, while construction aspect concerns on clarity of questions and options. Then the language or culture aspect assesses the communicative and effective language used to question items. The result of qualitative analysis is represented in Table 4.

Table 4. Scale of average score of qualitative analysis towards question items

No	Aspects	Scale of average score
A	**Content**	
1	Question items are appropriated with indicators	1.00
2	Content of questions is suitable with competencies	1.00
3	The options are homogen and logic	0.80
4	Each question has one correct answer	1.00
B	**Construction**	
1	Questions are formulated briefly, clearly and firmly	0.80
2	Formulation of questions and options contains statements that are required only	1.00
3	Questions do not describe the correct answer	1.00
4	Questions have none of negatively double statements	1.00
5	Options are homogen and logic based on content knowledge	1.00
6	Questions have functioning figures, graphs, diagrams, or others	0.90
7	Options relatively contain the same number of words	0.75
8	Options do not used “all options is correct or incorrect”	1.00

9	Options are arranged chronologically	0.80
10	Questions are independent from previous questions	1.00
C	**Language/culture**	
1	Questions use effective language	0.90
2	Questions use communicative language	0.95
3	Questions do not contain local language	1.00
4	Options do not repeat the same words, except for similar definition	0.90

Based on data presented in Table 4, the average of total score is 0.93 and it is classified as best valid eligibility. For each aspect, content (N-item= 4), construction (N-item=10), and language or culture (N-item=4) has average total score 0.95, 0.93, and 0,94 respectively. It means either average total score or average score per aspects have the best valid criteria. To support validity tests, reliability tests are carried out using statistical approaches for each aspect. Overall, Table 5 shows that the item score inside the aspect has positive correlation with total score. The reliability score of 0.94 proves excellent correlation, while construction aspect has strong correlation and language or culture shows reliable correlation (Taber, 2018). It is important to be noted that correlation above 0.7 meets expectation or acceptance (Yusup, 2018). Therefore, all question items used in this research are 100% valid and reliable.

Table 5. Reliability test of questions

No	Aspects	N-item	Reliability
1	Contents	4	0.94
2	Construction	10	0.92
3	Language/culture	4	0.89

Since the questions were totally valid and reliable, construction has the greater number of N-items in order to produce multiple-choice questions which have good quality. We must provide content and context without missing anything, such as lists of concepts, subject syllabi and specification tables. Besides that, it makes no ambiguity. Each content inside the questions must be clearly arranged or delimited, thus anyone would be able to understand it consistently or reliably at every condition (Moreno, et al., 2015). To ensure the result of measurement gained equal variances and covariances, it used a reliability test which is represented by Cronbach's alpha coefficient. This method allowed the measurement process to be reported as a confident result of evaluation due to its correlation between item score and total score (Bonett & Wright, 2014).

Correlation between difficulty index and discriminatory power

Discriminatory power is strongly related to difficulty index. High-performing students are able to answer the questions which have good discriminatory power and difficult or average difficulty index, whereas the low-performing students are able to answer the easy questions with poor discriminatory power (Khairani & Shamsuddin, 2016). The result of difficulty index and discriminatory power of questions is presented in Table 6.

Table 6. Results of difficulty index and discriminatory power

Question number	Difficulty index	Discriminatory power
1	Average	Good
2	Difficult	Excellent
3	Average	Excellent
4	Average	Good
5	Difficult	Excellent
6	Average	Good
7	Average	Excellent
8	Average	Excellent
9	Average	Excellent
10	Easy	Poor
11	Average	Good
12	Average	Excellent
13	Easy	Poor
14	Average	Good
15	Average	Excellent
16	Difficult	Excellent
17	Average	Excellent
18	Easy	Poor
19	Average	Excellent
20	Average	Excellent

Difficulty index is a percentage of students who answered the item test correctly (Patil, et al., 2016). Hence, the difficulty index shows the correlation with student achievement (Zainudin, et al., 2012). High performing students gained the greater total correct answer than low performing students. The performance of students is also represented by discriminatory power of question items. According to data presented in Table 6, it shows the positive correlation between difficulty index and discriminatory power. The correlation coefficient was calculated using Spearman-Brown equation (r=0,842). The coefficient between 0.84 - 0.90 is classified as reliable (Taber, 2018). This finding is better than previous research reported that the low correlation between discriminatory power and difficulty index is only 0.191 (Pande et al., 2013) and 0.195 (Sim & Rasiah, 2006).

It ranges from 0.00 to 1.00 (KoÇdar, 2016). The closer the index of the item test to 1.00, the easier that item is. Figure 1 describes that the questions included items with criteria that were easy (15%), average (70%) and difficult (15%). The highest percentage is presented by average questions. The questions with average criteria should be recorded in the questions bank, and then can be used again for other examinations. The previous study reported by Khairani & Shamsuddin (2016), there were 67% of items considered of good quality that will be kept for future testing. Besides that, the easy questions were generally not recommended for further tests. The difficult questions are proper to be used for the highest level examination, such as the olympiad.

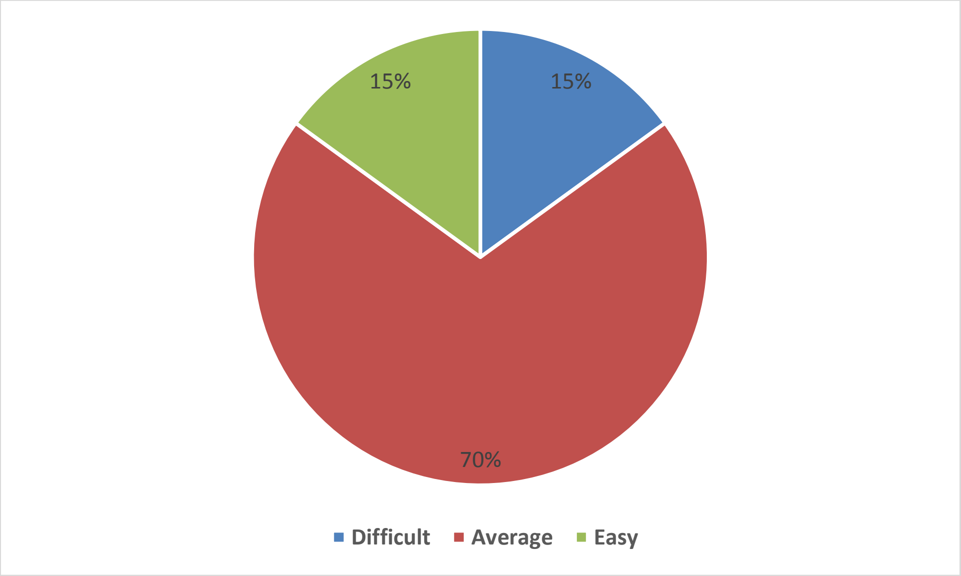

Figure 1. Percentage of Difficulty Index Criteria of Questions

Then, discriminatory power tests are done by presenting the final result as shown in Figure 2. There are 15% questions which have poor discriminatory power, and it means the questions did not distinguish high performing and low performing students. The rest are 85% questions distinguished students' performance in line with their level. Normally, the poor discriminatory power comes from the easy questions. Findings from a previous study showed that biology multiple-choice questions had average discriminatory power of 0.43 where more than 20% represented poor discriminatory power (Olutola, 2015). Another reported that average discriminatory power of multiple-choice questions is 0.22-0.38, or 52.58-69.5% of good discriminatory power (Chauhan, et al., 2013) Table 5 shows all three easy questions have poor discriminatory power, while average and difficult questions have good and excellent discriminatory power.

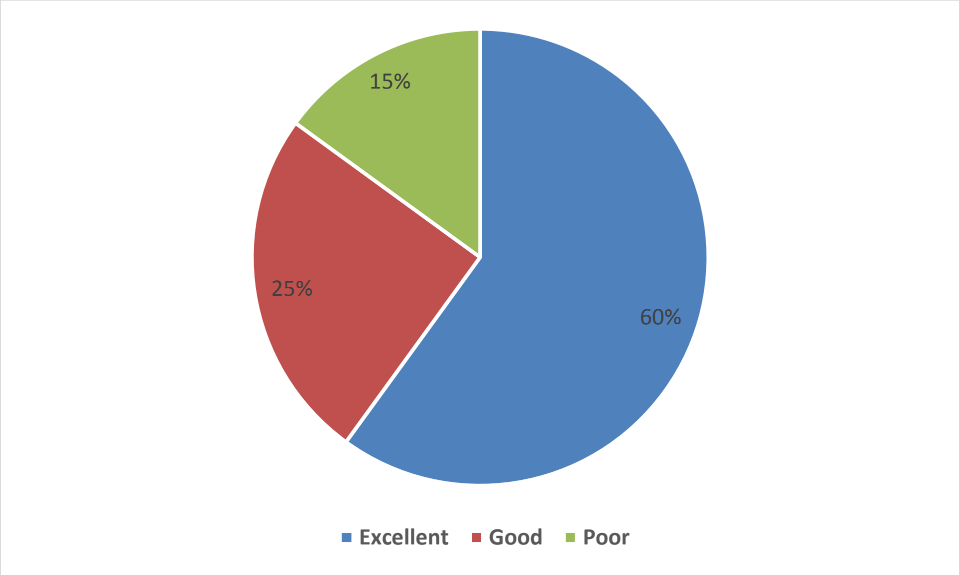

Figure 2. Percentage of Discriminatory Power of Questions

Functioning and nonfunctioning distractors

Multiple-choice questions will be objective when each question is equipped with several options or alternative answers. The alternative answers are three to five options among the options of each question, and one is the correct answer. Ideally, the questions must have 80 distractors distributed in 20 questions. It means each question has four functioning distractors. But, Table 9 shows that there are only nine questions with four functioning distractors. Then Table 7 shows that 16.25% are non-functioning distractors. Non-functioning distractors support low cognitive levels of students during answering the questions, so it must be removed from the questions in order to produce a better quality of questions. Previous study reported that there was an increment from 67 to 81% of functioning distractors after removal of non-functioning distractors (Ali & Ruit, 2015). Another explained that 64% were functioning distractors and 36% were non-functioning distractors (Shakurnia, 2019).

Table 7. Distractor function analysis of higher order thinking skills chemistry questions

No	Criteria	Number of distractors	Percentage of distractors
1	Functioning	67	83.75
2	Non-functioning	13	16.25

To evaluate functioning and non-functioning distractors of each item, it presented the further analysis using the following data in Table 8.

Table 8. Distractors distribution

Question number	The percentage of students who chose the options				
	A	B	C	D	E
1	5.56	0.00	66.67	16.67	11.11
2	11.11	16.67	55.56	5.56	5.56
3	22.22	50.00	27.78	0.00	0.00
4	5.56	38.89	22.22	16.67	16.67
5	5.56	38.89	22.22	27.78	5.56
6	16.67	5.56	50.00	27.78	0.00
7	22.22	27.78	38.89	5.56	5.56
8	22.22	16.67	38.89	11.11	16.67
9	0.00	44.44	22.22	22.22	5.56
10	5.56	11.11	83.33	0.00	0.00
11	22.22	16.67	27.78	27.78	5.56
12	33.33	5.56	22.22	38.89	0.00
13	11.11	0.00	72.22	11.11	5.56
14	33.33	27.78	33.33	5.56	0.00
15	16.67	22.22	44.44	5.56	11.11
16	16.67	55.56	27.78	0.00	0.00
17	0.00	5.56	50.00	11.11	33.33
18	11.11	0.00	11.11	72.22	5.56
19	5.56	38.89	11.11	33.33	11.11
20	5.56	5.56	16.67	11.11	61.11

Table 8 shows that several items have 0.00% non-functioning distractors such as question number 1, 3, 6, 9, 10, 12, 13, 14, 16, 17, and 18. Others research found that items with four-option that they reviewed had only one or two functioning distractors and none of five-option items had four functioning distractors (Haladyna & Downing, 1993). Another research done by Rahma, et al. (2017) explained that the average functioning distractors that applied for three and four options were 34 and 33%, respectively. Compared to this research, there are some percentages for the number of functioning distractors as shown as in Table 9.

Table 9. Percentage for number of functioning distractors

No	Number of functioning distractors	Amount of item	Percentage
1	4	9	45
2	3	8	40
3	2	3	15

According to Table 9, the question items had at least two functioning distractors. In addition, the biggest percentage is presented by four functioning distractors. Besides that, the three functioning distractors have similar percentages with four functioning distractors. This result is higher than previous research reported by Tarrant, et al. (2009) the study shows only 13.8% of all items had three functioning distractors and just over 70% had only one or two functioning distractors. Another reported that questions had 31.5 and 36.6% of two and three functioning distractors, respectively (Shakurnia, 2019).

Conclusion

According to the results of this research, the quantitative analysis towards higher order thinking skills chemistry questions is considered five aspects which are difficulty index, discriminatory power, validity, distractor function and reliability test. The questions are 100% valid with an average score of 0.93 as the best valid category. Further analysis using reliability tests shows that each aspect has positive correlation with Cronbach alpha of content, construction and language or culture 0.94; 0.92; and 0.89, respectively. Then, difficulty index and discriminatory power has correlation coefficient of 0.842, which means reliable. The final analysis for distractors show that overall questions have 83.75% having functioning distractors, while the remaining is non functioning distractors. The percentage is then divided into three groups; 45% of four functioning distractors, 40% or three functioning distractors and 15% stand for two functioning distractors.

References

Ali, S.H. & Ruit, K.G. 2015. The impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspective on Medical Education*, 4(1):244-251H.

Begum, T. 2012. A guideline on developing effective multiple choice questions and construction of single best answer format. *Journal of Bangladesh College of Physicians and* Surgeons, 30(2):159–166.

Bhat, S.K., & Prasad, K.H.L. 2021. Item analysis and optimizing multiple-choice questions for avisible question bank in ophthalmology: a cross-sectional study. *Indian Journal of Ophthalmology*, 69(2):343–346.

Bonett, D.G. & Wright, T.A. 2014. Cronbach's alpha reliability: interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, 36(1):3-15.

Butler, A.C. 2018. Multiple-choice testing in education: are the best practices for assessment also good for learning?. *Journal of Applied Research in Memory and Cognition*, 1(1):1–9.

Chauhan, Rameshbhai, P., Rathrod, Punjabhai, S., Rameshbhai, B., Rameshbhai, G., Ankit, Andvaryu, & Prahladbhai, A. 2013. Study of difficulty level and discriminating index of stem type multiple choice questions of anatomy in Rajkot. *Biomirror*, 4(6):37-40.

Dehnad, A., Nasser, H., & Hosseini, A.F. 2014. A comparison between three-and four-option multiple choice questions. *Procedia – Social and Behavioral Sciences*, 98(1):398–403.

DiBattista, D., Sinnige-Edger, J.-A., & Fortuna, G. 2014. The "none of the above" option in multiple-choice testing: an experimental study. *The Journal of Experimental Education*, 82(2):168–183.

Douglas, M., Wilson, J., & Ennis, S. 2012. Multiple-choice question tests: a convenient, flexible and effective learning tool? a case study. *Innovation in Education and Teaching International*, 49(2):111-121.

Griethuijsen, R.A.L.F., Eijick, M.W., Haste, H., Brok, P.J., Skinner, N.C., & Mansour, N. 2014. global patterns in students' view of science and interest in science. *Research in Science Education,* 45(4):581–603.

Hadisaputra, S., Ihsan, M.S., Gunawan, & Ramdani, A. 2020. The development of chemistry learning device based blended learning model to promote students' critical thinking skills. *Journal of Physics: Conference Series*, 1521(1):1–5.

Haladyna, T.M., & Downing, A.M. 1993. How many options is enough for a multiple-choice test item?. *Educational and Psychological Measurement,* 53(1):999–1010.

Hamdi, S., Suganda, I.A., & Hayati, N. (2018). Developing Higher-Order Thinking Skills (HOTS) Test instrument using lombok local cultures as contexts for junior secondary school mathematics. *Research and Evaluation in Education*, 4(2):126–135.

Harta, J., Rasuh, N.T., & Seriang, A. (2020). using hots-based chemistry national exam questions to map the analytical abilities of senior high school students. *Journal Science Learning*, 3(3):143–148.

Ichsan, I.Z., Sigit, A.V., Miarsyah, M., Ali, A., Arif, W.P., & Prayitno, T.A. 2019. HOTS-AEP: Higher order thinking skills from elementary to master students in environmental learning. *European Journal of Educational Research*, 8(4):935–942.

Isa, M., Khaldun, I., & Halim, A. (2017). Penerapan model pembelajaran kooperatif tipe tai untuk meningkatkan penguasaan konsep dan berpikir tingkat kritis siswa pada materi hidrokarbon. *Jurnal IPA dan Pembelajaran IPA,* 1(2):213-223.

Khairani, A.Z., & Shamsuddin, H. 2016. Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. *Assessment for Learning Within and Beyond the Classroom*, 1(1):417-426.

Kingston, N.M., Tiemann, G.C., Miller-Jr, H.L., & Foster, D. 2012. An analysis of the discrete-option multiple choice item type. *Psychological Test and Assessment Modelling*, 54(1):3–19.

KoÇdar, S., Karadag, N., & Sahin, M.D. 2016. Analysis of the difficulty and discrimination indices of multiple-choice qustions according to cognitive levels in an open and distance learning context. *The Turkish Online Journal of Educational Technology*, 15(4):16–24.

Mahjabeen, W., Alam, S., Hassan, U., Zafar, T., Butt, R., Konain, S., & Rizvi, M. 2014. difficulty index, discrimination index and distractor efficiency in multiple choice questions. *An Official Journal of Shaheed Zulfiqar Ali Bhutto Medical University (SZABMU)*, 13(4):310-316.

Moreno, R., Martinez, R.J., & Muniz, J. 2015. guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, 27(4):388-394.

Olutola, A.T. 2015. Item difficulty and discrimination indices of multiple choice biology tests. *Liceo Journal of Higher Education Research*, 11(1):16-30.

Osundare, A.G., & Omirin, M.A. (2016). Determining the differences in the difficulty and discriminating indices of chemistry completion and matching test formats. *European Journal of Research and Reflection in Educational Science*, 4(2):65–70.

Pande, S.S., Pande, S.R., Parate, V.R., Nikam, A.P., & Agrekar, S.H. 2013. Correlation between difficulty & discrimination indices of mcqs in formative exam in physiology. *South-East Asian Journal of Medical Education,* 7(1):45–50.

Patil, R., Palve, S.B., Vell, K., & Boratne, A.V. 2016. Evaluation of multiple choice questions by item analysis in a medical college at Pondicherry, India. *International Journal of Community Medicine and Public Health*, 3(6):1612–1616.

Puspitasari, Y.D., & Nugroho, P.A. 2020. Peningkatan *higher order thinking skill* dan kemampuan kognitif pada mahasiswa melalui pendekatan *science, environment, technology and society* berbantuan modul pembelajaran. *JIPI (Jurnal IPA dan Pembelajaran IPA)*, 4(1):11-28.

Rahma, A., Shamad, M., Idris, M.E.A., Elfaki, O., Elfakey, W., & Salih, K.M.A. 2017. Comparison in the quality of distractors in three and four options type of multiple choice questions. *Advances in Medical Education and Practice*, 8(1):287-291.

Rahmati, Yusrizal, & Hasan, M. 2017. Critical thinking skills enhancement of students through inquiry learning model laboratory based on reflection of the light. *JIPI (Jurnal IPA dan Pembelajaran IPA)*, 1(1):34-41.

Shakurnia, A. 2019. A Survey on distractors in multiple-choice questions and its relationship on difficulty and discriminative indices. *Iranian Journal of Medical Education*, 19(1):180-188.

Sim, SM., & Rasiah, R.I. 2006. Relationship between item difficulty and discrimination indices in true/false-type multiple choice questions of a pra-clinical multidisciplinary paper. *Ann Acad Med Singapore*, 35(1):67–71.

Sinurat, R., Nevrita, & Hindrasti, N.E.K. 2020. Identifikasi tingkat kemampuan berpikir kritis siswa pada materi asi eksklusif dan program keluarga berencana. *JIPI (Jurnal IPA dan Pembelajaran IPA),* 4(1):60-69.

Sugiyono. 2009. *Metode penelitian pendidikan (pendekatan kuantitatif, kualitatif, dan R&D*. Bandung: CV. Alfabeta.

Suwono, H., Pratiwi, H.E., & Susilo, H. 2017. Enhancement of student's biological literacy and critical thinking of biology through socio-biological case-based learning. *JIPI (Jurnal IPA dan Pembelajaran IPA),* 6(2):213-220.

Taber, K.S. 2018. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48(1):1273–1296.

Taib, F., & Yusoff, M.S.B. 2014. Difficulty index, discrimination index, sensitivity, and specificity of long case and multiple choice questions to predict medical students' examination performance. *Journal of Taibah University Medical Science*, 9(2):110-114.

Tarrant, M., Ware, J., & Mohammed, A.M. 2009. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *BMC Medical Education*, 9(40):1–8.

Towns, M.H. 2004. Guide to developing high quality, reliable, valid multiple-choice assessments. *Journal of Chemical Education*, 91(1):1426–1431.

Tsaparlis, G. 2020. Higher and lower-order thinking skills: the case of chemistry revisited. *Journal of Baltic Science Education*, 19(3):467–483.

Yusup, F. 2018. Uji validitas dan reliabilitas instrumen penelitian kuantitatif. *Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan*, 1(1):17–23.

Zainudin, S., Ahmad, K., Ali, N.M., & Zainal, F.A. 2012. Determining course outcomes achievement through examination difficulty index measurement. *Procedia – Social and Behavioral Sciences*, 59(1):270-276.