JKPK (JURNAL KIMIA DAN PENDIDIKAN KIMIA), Vol. 6, No. 3, 2021 pp. 343-351
Chemistry Education Study Program, Universitas Sebelas Maret ISSN 2503-4146
https://jurnal.uns.ac.id/jkpk ISSN 2503-4154 (online)

ITEM QUALITY ANALYSIS OF CHEMISTRY FINAL SEMESTER TEST IN THE ACADEMIC YEAR OF 2017/2018, 2018/2019 AND 2019/2020

Nazatin Nuroini, Suryadi Budi Utomo*, and Sulistyo Saputro

Chemistry Education Study Program, Faculty of Teacher Training and Education, Universitas Sebelas Maret
Jl. Ir. Sutami No.36, Surakarta City, Central Java, 57126, Indonesia

*Correspondance, e-mail: sbukim98@staff.uns.ac.id

Received: September 12. 2021 Accepted: December 18, 2021 Online Published: December 28 , 2021

DOI : 10.20961/jkpk.v6i3.54999

ABSTRACT

Item analysis is a process in which both students' answers and test questions are evaluated in order to determine the quality of the items and the test as a whole in the standardized and objective evaluation of student performances. Evaluation is needed to define how much the participants' learning outcomes have changed from their beginning abilities to their abilities after completing the educational process. This research examines the items' quality from a quantitative standpoint. The aims of this study are to determine the difficulty index, item discrimination, distractor effectiveness, and reliability of the final semester test for Chemistry Subject Class X MIPA SMAN 1 Wonosegoro, Boyolali Regency. The research was carried out at SMAN 1 Wonosegoro, Boyolali Regency, using a quantitative descriptive technique. This study's population consists of response data from all 212 students in class X MIPA Chemistry throughout 2019/2020, 2018/2019 and 2017/2018 academic years. Documentation techniques are used to collect data. The data was analyzed quantitatively with the ANATES 4.0.9 version. According to the findings, in three consecutive academic years, the difficulty index was medium means is good because it is neither too complex nor too simple. Item discrimination is acceptable and meets the standards of sufficient, good, and exceptional. The distractor effectiveness was functions and the reliability value in the Academic Year of 2017/2018 was sufficient at 0.45, but it was high at 0.62 and 0.78 in the Academic Year of 2018/2019 and 2019/2020. The finding of this study item analysis is a crucial process in creating tests. This is about the impact of the accuracy of students' scores on test quality.

Keywords: *Items Analysis, Chemistry, SMAN 1 Wonosegoro, ANATES 4.0.9.*

INTRODUCTION

Teachers in Indonesia are given the responsibility of evaluating the quality of the question instrument for students by asking questions or implementing acceptable, rational, and scientific policies based on criterion standards. In Ghana, the same thing happened. The teacher is given a central role in the testing and evaluation process [1]. Italy is also working on a significant project in Genoa to improve the skills of teachers from several primary and secondary schools to apply standard and objective student achievement evaluations in a systematic way [2].

This indicates that learning evaluation is a crucial activity in the education system. Because of its importance, evaluation has

become an intrinsic aspect of the learning process [3]. As a result, teachers must be well-versed in testing techniques to evaluate students' development in a reliable and valid way [4]. The purpose is to measure and identify the level of change that occurs in order to achieve specified goals and to know the quality of the test [5].

A good quality test must have well-constructed items that enable teachers to accurately measure students' abilities [6]. At least three criteria should be included, encompassing practicality, reliability, and validity [7]. The operating budget, time limit, implementation, and test scoring system are all examples of practicality [7]. In terms of reliability, the test result should provide consistent results under a variety of conditions [8]. While the validity of a test is the amount to which it measures only what it is designed to measure [9]. So, reliability refers to dependability and validity refers to the ability of a test to measure what should be measured in relation to the learning goals or competences to be achieved [10].

In the process of implementing evaluation activities, there are two types of testing techniques: testing and non-testing. The testing technique is the most commonly employed by schools in Indonesia to determine the success of the teaching and learning process. Generally, Indonesian teachers apply tests, specifically called summative tests, to evaluate students at the end of the learning process.

To find out the quality of the items, the teacher must first analyze the items. A well-test helps the teacher to evaluate students' understanding of specific content delivered in class in an accurate and systematic way [4]. Item analysis is the process of collecting, summarizing, and analyzing data from students responses in order to evaluate the test items quality and determine whether or not items are of sufficient quality to include in a test [4]. An item analysis enables identifying the quality of Multiple-Choice Questions (MCQs) based on reliability, validity, Difficulty Index (DI), Item Discrimination (ID) and Distractor Effectiveness (DE) [11].

(MCQs) is the most commonly used tool for assessing the knowledge capabilities of students. Meanwhile, the first part of an MCQ is called the stem, and it contains the problem or question while the second part is called the response and it contains a list of potential explanations [12]. Using a benchmark modified to the opinion Likert, or popularly known as the scale Likert, interpret the effectiveness of the distractors based on these results are if the four answers to the question fraud work properly, the question has a very good distractor effectiveness criterion. If there are three distractors that work well, it is good. If there are two distractors that work well, it is enough. If only one item distractor works well, the problem-distracting. The last one, if the four distractors do not function well, is not good [13].

The difficulty index is one of the question tools. The percentage of students who think an item is easy or difficult is known as the difficulty index [14]. The difficulty index is better known by the symbol P, which stands for "proportion" [15]. A good item that can also be seen from its item discrimination is the ability to differentiate between high-ability and low-ability students and ranges between 0 and

1 [13]. An item test can reach its ideal index when high achievers answer correctly more often than low achievers [16]. And if the value of D is negative, everything is not good. So items that say a negative D value should be discarded [17].

SMAN 1 Wonosegoro was chosen to be the object of research because neither teacher nor other researchers had analyzed the items. This analytical activity is required to determine the students' ability to understand the subject over a specified period of time. Because providing evaluations on the final semester, test is required to determine students' capacity to master the content and track their learning progress. Item analysis is one of the methods for determining the quality of the test and can be used as a factor in determining whether or not any of the questions will be used in future periods.

METHOD

The design in this study uses a quantitative approach and uses quantitative descriptive methodologies. Quantitative research is conducted by collecting information or using a structured list of questions to generate quantitative data in the form of numbers. The results of this study were determined by analyzing the questions in terms of difficulty index, item discrimination, distractor effectiveness, and reliability then were analyzed by using ANATES 4.0.9 version.

The index difficulty is determining the percentage of students who correctly answer [18]. This definition is comparable to [19], who states that "difficulty index refers to the percentage of students who think an item is easy or difficult." The index of difficulty has three ranges to interpret, resulting in the value of index difficulty are as follows [8].

Table 1. Classification of index difficulty.

No	P	Criteria
1	0.0 – 0.3	Difficult
2	0.3 – 0.7	Sufficient
3	0.7 – 1.0	Easy

Item discrimination represents a question test's capacity to distinguish between outstanding and non-outstanding students [18]. For the item discrimination, there are five criteria of interpretation for the resulting numbers are as follows [13].

Table 2. Classification of item discrimination

No	ID	Criteria
1	<0	Very Bad
2	0.0 – 0.2	Bad
3	0.2 – 0.35	Medium
4	0.35 – 0.7	Good
5	0.7 – 1.0	Very Good

Another characteristic of item analysis is distractors. Only multiple-choice exams can be used to analyze this characteristic. At least 5% of the respondents must choose a distractor, especially those who are low achievers [20]. To find out the reliability of the test were analyzed for three aspects: substance, construction, and language or culture. Cronbach alpha was used to quantitatively analyze the relationship between item score and total score.

In this study, the objective of quantitative descriptive research is to explain a condition that will be researched with the help of another literature study, which can

strengthen researchers' assumptions in the data analysis process based on the current state of the object and in drawing a conclusion, so that the item can be determined as a matter of good quality or vice versa. The data was collected in the form of text items for the final semester test of Chemistry Subjects for Class X MIPA at SMAN 1 Wonosegoro, Boyolali Regency, on three sets of questions that were tested in different years with different students for three consecutive academic years, namely 2019/2020, 2018/2019 and 2017/2018.

RESULTS AND DISCUSSIONS

Table 3. The amount of MCQ in Chemistry for students in grades X MIPA at SMAN 1 Wonosegoro, Boyolali Regency.

Academic Year	Amount of Students	Amount of Item
2017/2018	57	40
2018/2019	63	35
2019/2020	92	40

According to Table 3, 212 respondents were collected from all class X MIPA SMAN 1 Wonosegoro, Boyolali Regency, throughout the course of 3 academic years.

The quantitative analysis carried out in this study aims to determine difficulty index, item discrimination, distractor effectiveness, and reliability of the items in the Class X MIPA subject at SMAN1 Wonosegoro, Boyolali Regency. In the current study, in the academic years 2017/2018 and 2019/2020, there were 40 questions. While there were 35 questions in 2018/2019, but only 34 could be used because 1 question didn't have an answer key.

Table 4. Classification of difficulty index

Criteria	Academic year		
	2017/ 2018 (%)	2018/ 2019 (%)	2019/ 2020 (%)
Easy	40	38.1	5
Sufficient	60	58.8	67.5
Difficult	0	2.9	27.5

From Table 4 above, the majority of the 40 test items are acceptable in terms of index difficulty. In the academic year of 2017/2018, 16 items (40%) include easy criteria, 24 items (60%) included sufficient criteria. In the Academic Year of 2018/2019, accounting for 13 items (38.1%) include easy criteria. 20 items (58.8%) include sufficient criteria and 1 item (2.9%) include difficult criteria. And in the academic year of 2019/2020, 2 items (5%) namely in easy criteria, 27 items (67.5%) is most sufficient criterion and 11 items (27.5%) with difficult criteria.

There should be 25% easy items, 50% sufficient items, and 25% difficult index in the ideal test [21]. According to the findings, the test package does not have a proportional index difficulty. A well-constructed item cannot be too easy or difficult [19]. However, the difficulty index on the test must be dominated by sufficient criteria. If the questions are getting more difficult, there will be fewer test-takers who correctly answer the questions. So, a good question should be a medium difficulty index.

These findings are in accordance with those of other research e.g. Karimah et al., [29], Reza et al., [31], Hasanah et al., [32] have attempted to find out index difficulty of chemistry lessons for senior high school were more sufficient than other categories items. Nevertheless, the portion among easy,

sufficient, and difficult indexes is not balanced. These results could be explained. Cognitive variables can influence the index of difficulty [22]. Comprehending, coding, transitioning, scrutinizing, and working memory are all cognitive variables that influence index difficulty measurement [23].

In addition to the index of difficulty, there are several categories for quality assessment of the item analysis, as shown in table 3 below this. The item discrimination of these questions was included in the bad criteria for 2 consecutive academic years, 2017/2018 and 2018/2019. For the next academic year 2019/2020, the item discrimination of questions was included in the good criteria.

Table 5. Classification of item discrimination.

Criteria	Academic Year		
	2017/ 2018 (%)	2018/ 2019 (%)	2019/ 2020 (%)
Very Bad	12.5	8.8	12.5
Bad	32.5	32.3	25
Medium	27.5	32.3	20
Good	22.5	20,5	35
Very Good	5	5.9	7.5

From the Table 5 above, the item discrimination in the academic years of 2017/2018 have D value of 22 items (55%) however, 11 items (27.5%) of them should be revised. For the Academic Year of 2018/2019 is 20 items (58.7%), which means the "D" is good, but there needs to be a revision in the items with medium criteria in 11 items (32.3%). And for the item discrimination in the academic year of 2019/2020 has a proportion of 25 items (62.5%) is good, however 8 items (20%) of them should be revised first. It must be revised for items with bad criteria, and items of very bad criteria must be eliminated.

When viewed from the pattern of student answers in the 2017/2018 and 2018/2019 academic years, there are several factors that affect the results of item discrimination each year because there are some test takers who are uncertain about the answer because there are several empty and multiple answers to questions. Meanwhile, in the academic year of 2019/2020, the percentage of item discrimination included in the good criterion. Questions like these are ideal because they can identify students' talents [16]. Hence, the majority of the items are accepted and can be used as an item bank. There are a few things that should be revised.

Negative “D” (very bad criteria), simply means that students of lower ability guess at the correct answer without really understanding it [24]. In the current study, there are some very bad results criteria in three academic years in a row. Items with negative D decrease the test's validity and should be eliminated from the question group [24]. The most likely explanation was a wrong key, confusing question framing, or generalized poor student preparation [13].

These findings in the academic years of 2017/2018 and 2018/2019 are comparable to other studies e.g. Maharani [28], Karimah et al., [29], Manfenrius et al., [30] presented more than half of the items in this study qualified as good discrimination indexes.

The Table 6 below, the quality of the distractor in three consecutive academic years (2017/2018; 2018/2019; 2019/2020) is

effective with the percentage each year in a row is 26 items (65%), 22 items (64.7%), 36 items (90%).

Table 6. Classification of distractor effectiveness

Criteria	Academic Year		
	2017/ 2018 (%)	2018/ 2019 (%)	2019/ 2020 (%)
Not Good	15	8.8	2.5
Poor	20	26.5	7.5
Enough	30	23.5	27.5
Good	15	11.8	27.5
Very Good	20	29.4	35

Criteria for the interpretation of the distractor effectiveness a question, namely, 0 means not good; 1 means poor; 2 is enough; 3 means good, and 4 means very good," writes by [25]. A good question, according to this statement, has at least two distractors. Distractors with enough, good, and very good criteria can be used again in future tests, however questions with poor and not good criteria are rejected, meaning that the question should be corrected first or replaced with another distractor. The functioning of the distractors influences the level of difficulty of the question because if one or two distractors are not working effectively, the difficult index value will decrease (towards difficult/ extremely difficult) since the chance of students answering correctly decreases [27].

According to Maharani [28] research, distractors are effective in 80%. This relates to the 92.5% on the index difficulty with moderate criteria. A comparable result was found in Hartati [33], with distractors being effective of 53% with index difficulty being 50%. In contrast, Iskandar [9] found that the distractor function value of 11.1% because the index difficulty of the difficult criteria dominated as much as 57.5%. The index difficulty of the question is influenced by the functioning of the distractors because if one or two distractors are not working effectively, the difficult index value will decrease (towards difficult/extremely difficult) since the chance of students answering correctly decreases [9].

Table 7. Results of Reliability Test

Academic Year	Value Reliability (r^2)	Interpretation
2017/2018	0.45	Low
2018/2019	0.62	Low
2019/2020	0.78	High

The reliability coefficient, which can be found in Table 7 below, was 0.45 in the Academic Year of 2017/2018 and 0.62 in the academic year of 2018/2019 is low or unreliable meaning that it should not be used in future exams. The reliability coefficient of 0.78 in 2019/2020 is a high criterion meaning that the questions are reliable and can be used in exams. With 5 criteria, the reliability coefficient ranges from 0 to 1. If the test results show the construct reliability is 0.7, the items are said to have a reliable construct [26].

CONCLUSION

The ANATES 4.0.9 version results provide evidence that generally, these items are good items. However, some items had to be revised and eliminated or replaced with other items and could be used in tests or saved for future testing. These poor items may influence the students' level of

understanding and challenging materials or topics, which may have an impact on the difficulty index. Item discrimination and distractor effectiveness can be influenced by ambiguity in the options or even the key answer, ambiguity of instructions and the number of question items and participants analyzed. Although the fact that this study has some advantages, it also has some limitations, such as the study's limited variables and data. These findings may be useful to teachers or test creators as advice for making changes to the way they create test items. Various researchers should add other techniques of analyzing test items to compare the outcomes in future studies. Other researchers should finish the study with qualitative analysis to get more in-depth results.

REFERENCES

[1] J. Anamuah-Mensah & K. A. Quaigrain, "Teacher competence in the use of essay test," *The Oguaa Educator University of Cape Coast,* vol. 12, pp. 31–42, 1998.

[2] A. Siri & F. Michela, "The use of item analysis for the improvement of objective examination," *Procedia – Social and Behavioral Sciences*, vol. 29, pp. 188-197, 2011. DOI: 10.1016/j.sbspro.2011.11.224

[3] M. Ramlawati., M. Anwar., S. R. Yunus., & M. Nuswowati, "Analysis of Students' Competence in Chemistry Cognitive Test Construction Based on Revised Bloom's Taxonomy," *Journal of Physics: Conference Series*, vol. 1567, pp. 1-6, 2020. DOI:10.1088/1742-6596/1567/4/042006

[4] Q. Kennedy & A. K. Arhin, "Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation," *Cogent Education,* vol. 4, 2017. DOI:10.1080/2331186X.2017.1301013

[5] W. R. Tyler, Basic Principle of Curriculum and Instruction, Chicago: The University of Chicago Press, 1949.

[6] P. Jefri., Mujassam, & Y. T. Alberto. "Item Analysis Using Rasch Model in Semester Final Exam Evaluation Study Subject in Physics Class X TKJ SMK Negeri 2 Manokwari," *Physics Educational Journal*, vol. 1(1), pp. 43-51, 2018. DOI: 10.37891/kpej.v1i1.40

[7] H. D. Brown, Teaching by Principles: An Interactive Approach to Language Pedagogy (2nd ed.), New York: Longman, 2001.

[8] G. Flucher & F. Davidson, Language Testing and Assessment: An Advance Resource Book, Routledge, 2007.

[9] J. B. Heaton, Writing English Language Test, New York: Longman, 1975.

[10] V. M. Amalia, "Item Analysis of English Final Semester Test," Indonesian *Journal of EFL and Linguistics*, vol. 5, no. 2, pp. 491-504, 2020. DOI: 10.21462/ijefl.v5i2.302

[11] M. L. Hidayat, W. H. Prasetiyo, & J. Wantoro, "Pre-service student teachers' perception of using google classroom in a blended course," *Humanities and Social Sciences Reviews*, 7(2), 363–368, 2019. DOI: 10.18510/hssr.2019.7242

[12] M. S. Velou & E. Ahila, "Refine the multiple choice questions tool with item analysis," *International Archives of Integrated Medicine*, vol. 7(8), pp. 80-85, 2020.

[13] G. Mehta and V. Mokhasi, "Item Analysis of Multiple-Choice Questions-An Assessment of the Assessment Tool". *International Journal of Health Science and Research*, vol. 4 (7), pp. 197-202, 2014.

[14] N. N. Agu, C. Onyekuba C, & A. C. Anyichie, "Measuring teachers' competencies in constructing classroom-based tests in Nigerian secondary schools: Need for a test construction skill inventory", *Educ. Res. Rev*. vol. 8, no. 8, pp. 431-439, 2013.
DOI: 10.5897/ERR12.219

[15] A. S. Ingale, P. A. Giri, & M. K. Doibale, "Study on item and test analysis of multiple-choice questions amongst undergraduate medical students," *International Journal of Community Medicine and Public Health*, vol. 4, no. 5, pp. 1562-1565, 2017.
DOI:10.18203/2394-6040.ijcmph20171764

[16] M. R. Hingorjo, & F. Jaleel, "Analysis of One-Best MCQs: The Difficulty Index, Discrimination Index and Distractor Efficiency," *JPMA-Journal of the Pakistan Medical Association*, vol. 62(2), pp. 142–147, 2012.

[17] S. Arikunto, Dasar-dasar Evaluasi Pendidikan, Jakarta: Bumi Aksara, 2013.

[18] T. M. Haladyna, *Developing and Validating Multiple-Choice Test Items* (3rd ed.). Lawrence Erlbaum Associates Publisher, 2004.

[19] A. Iskandar, & M. Rizal, "Analisis Kualitas Soal Perguruan Tinggi Berbasis Aplikasi TAP," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 21, no. 2, pp. 12-23, 2017.
DOI: 10.21831/pep.v22i1.15609

[20] D. Rosana, & D. Setyawarno, "Statistik Terapan Untuk Penelitian Pendidikan," UNY Press, 2017.

[21] Kunandar, *Penilaian Autentik: (Penilaian Hasil Belajar Peserta Didik Kurikulum 2013).* Raja Grafindo Persada, 2013.

[22] P. J. Sung, S. W. Lin, & P. H. Hung, "Factors Affecting Item Difficulty in English Listening Comprehension Tests," *Universal Journal of Educational Research*, *3*(7), 451–459, 2015. DOI: 10.13189/ujer.2015.030704

[23] E. Danili & N. Reid, "Cognitive Factors That Can Potentially Affect Pupils' Test Performance," *Chemistry Education Research and Practice*, *7*(2), 64–83, 2006.
DOI: 10.1039/B5RP90016F

[24] S. Gajjar, R. Sharma, P. Kumar, & M. Rana, "Item and test analysis to identify quality Multiple Choice Questions (MCQs) from an assessment of medical students of Ahemdabad, Gujarat," *Indian journal of Community Medicine*, vol. 39, pp. 17-20, 2014.
DOI: 10.4103/0970-0218.126347

[25] Sugiyono, *Metode Penelitian Pendidikan: Pendekatan Kuantitatif, Kualitatif, Dan R&D*, Bandung: Alfabeta, 2010.

[26] S. Bintarti & E. N. Kurniawan, "A study of revisit intention: Experiential quality and image of Muara Beting tourism site in Bekasi District". *European Research Studies Journal*, 20(2), 521–537, 2017.
DOI: 10.35808/ersj/657

[27] D. Kheyami, A. Jaradat, T. Al-Shibani, & F. A. Ali, "Item Analysis of Multiple-Choice Questions at the Department of Paediatrics," Arabian Gulf University, Manama, Bahrain. *Sultan Qaboos Univesity Medical Journal*, *18*(1), 2018.
DOI: 10.18295/squmj.2018.18.01.011

[28] A. V. Maharani, "Item Analysis of English Final Semester Test" *Indonesian Journal of EFL and Linguistics,* vol. 5 (2): 491-504, 2020.
DOI: 10.21462/ijefl.v5i2.302

[29] U. Karimah, H. Retnawati, D. Hadiana, Pujiastuti, & E. Yusron, "The characteristics of chemistry test items on nationally-standardized school examination in Yogyakarta City," *Research and Evaluation in Education*, vol. 7 (1), pp. 1-12, 2021.
DOI: 10.21831/reid.v7i1.31297

[30] A. Manfenrius, G. Sutapa, & B. Wijaya, "Item Analysis on English Summative Test at The Eighth Grade Junior High Schools in Pontianak," *Jurnal Pendidikan Dan Pembelajaran Khatulistiwa*, *4*(12), 1–10, 2015.

[31] M. Reza, K. Puspita, & C. Oktaviani, "Quantitative Analysis Towards Higher Order Thinking Skill of Chemistry Multiple choice Questions for University Admission," *Jurnal IPA dan Pembelajaran IPA*, vol. 5(2): 172-185, 2021. DOI: 10.24815/jipi.v5i2.20508

[32] I. Hasanah, J. Copriady, & A. Thaib, "Analisis Butir Soal Ujian Semester Ganjil Pelajaran Kimia Kelas XI IPA SMA Negeri 10 Pekanbaru Tahun Pelajaran 2013/2014," *JOM,* vol. 2(1), 1-10. 2015.

[33] N. Hartati, H. P. S. Yogi, "Item Analysis for a Better-Quality Test," *ELIF,* vol. 2(1), 59-70, 2019. DOI: 10.24853/elif.2.1.59-70