Analysis of Multiple Choice Questions for Japanese Language Mid-Term Exam at Public Senior High School

Rizki A. N. Dayanto(✉), Syamsul Sodiq, and Miftachul Amri

Universitas Negeri Surabaya, Surabaya, Indonesia
rizki.21006@mhs.unesa.ac.id

Abstract. Evaluation is a process of finding information related to student outcomes. This study aims to analyze the quality of school middle test items in the form of multiple choice conducted by class XI students of Public Senior High School 1 Purwoasri. Quantitative descriptive research is used to determine and describe the test quality, such as validity, reliability, and difficulty level. Calculation of validity, reliability, and level of difficulty of the questions is done using a computer program (ANATES). The result of the analysis showed that from a total of 40 items, only 2 items were valid. The reliability coefficient obtained is 0.08, which means the question item is unreliable since it is far from the reliable limit of 0.60. Furthermore, the classification of the difficulty level obtained 32 easy questions, 6 medium questions, and 2 difficult questions. From the results above, it is necessary to improve the items since they have not been achieved from the minimum standard of item assessment.

Keywords: Multiple choice · Validity · Reliability · Level of difficulty · ANATES

1 Introduction

Education is an effort to make students active in their lives and ready to face the future. Success in terms of education will make the nation advanced and have the quality to become the foundation of a country. For this reason, it is necessary to develop quality in the learning process or improve education quality. In carrying out education, teaching methods are needed. Then, to find out whether the method can achieve educational goals, an evaluation is needed.

Evaluation is a process carried out to measure and assess, which can also mean collecting information in the form of data following the planned objectives. The results of the evaluation are also data sources for obtaining information to show whether the teaching objectives have been achieved or not. The results obtained can be used as teacher feedback and determine learning effectiveness. The feedback is used to improve the learning program [1]. As an evaluator, the teacher plays a very important role in the success of his students. Changes that occur in students both behavior and knowledge

S. Setiawan et al. (Eds.): IJCAH 2022, ASSEHR 724, pp. 1791–1801, 2023.
https://doi.org/10.2991/978-2-38476-008-4_194

are the contribution and encouragement made by the teacher. Therefore, teachers need to give an objective assessment.

Evaluation is a systematic activity used to determine student progress during the learning process [2]. By collecting data and comparing it with specific criteria, the teacher can understand the abilities and how the progress of students. The data collected can be the students' behavior or appearance during the learning process. Besides, assessment is an activity to obtain facts using critical and careful steps. It is carried out since there are problems that require correct answers, such as why student achievement is low [3]. Assessment can also be used later for grade promotion or graduation decisions. Several evaluation techniques and several tests are used in its implementation.

A test is an evaluation tool that consists of subjective and objective tests. Subjective tests are essays or descriptions, while objective tests are matching, true-false, and multiple-choice tests [4]. Multiple choice test questions are questions for which answer options have been provided. Multiple-choice questions contain problems that will become the main questions and possible answer choices [5]. Objective assessment and easy examination are the reasons for the frequent use of multiple-choice tests. Another advantage of multiple-choice questions is high reliability and can also be used to measure the level of thinking ability [6].

Tests can be used to obtain information on the success or failure of students in achieving the objectives contained in the curriculum [7]. The tests can be in the form of questions or assignments. For example, the tests are conducted during the mid-term and final exams. To determine the quality, the questions used as tests must be analyzed first and meet the criteria of being valid, reliable, objective, practical, and economical [8]. Not only the results, but the teacher must also provide an assessment of the test questions used.

Questions that provide information about students' abilities and developments can be considered quality questions. To gain quality items, item analysis is needed. The steps must be taken to determine the test quality level, and the items are called test quality analysis [9]. Item analysis is an assessment of test questions carried out to obtain a quality set of questions. Another goal is to identify which questions are good, less good, and not good so that improvements can be made. The systematic procedure of problem analysis will provide exact information on the test items compiled [8].

Validity is the accuracy level between the data listed by the researcher and on the research object. So, valid data is the same between the data listed by the researcher and in the research object [10]. Validity comes from the word validity which means the accuracy level and accuracy in measurement. If the test measures what it is supposed to measure, it can be said to be valid. For example, in the imagery, "speedometer is a valid measuring instrument to measure the object speed, but is not valid if used to measure weight". In education, a knowledge test of a field of study is not a valid measuring tool that can be used to measure attitudes towards the subject area. An item is said to have high validity if there is a match between the item and the total score, or other expressions, namely the question score and the total score, have significant positive relevance. Validity aims to determine how accurate the measuring instrument is in carrying out its measuring function so that the data obtained can be consistent with the research objectives.

Reliability is the determination level of the measurement of an object. When a test can give a fixed result, it can be said that it has a high confidence level. Test reliability is related to the problem of determining test results, or if the test results change, then the changes can be considered meaningless. Based on the explanation above, it can be concluded that reliability is the confidence level in the tests used to measure student learning outcomes [11]. If the reliability is low, the test can make students hesitate to answer the questions. Meanwhile, if the reliability is high, the test has reliable and consistent measurement results.

Every semester in the school, a mid-term exam is held in the middle of the semester. The purpose of implementing the mid-term exam is to measure the achievement of student competencies after undergoing learning in half a semester. In addition, by knowing the results of the mid-term exam scores, educators can make learning improvements for the next half-semester. The test is a multiple-choice objective test with five answer choices. The items tested represent the subject matter and teaching objectives.

Learning Japanese has a different difficulty level than other languages since learning how to read and write letters (katakana, hiragana, and kanji) is also necessary. Thus, this study aims to analyze the quality of mid-term exam items conducted by XI MIPA 4 students of Public Senior High School 1 Purwoasri. To determine the test quality, it is necessary to test the validity, reliability, and difficulty level using ANATES. It is software used to analyze items, which was developed by a psychology lecturer at UPI and a computer consultant. Things that can be used in ANATES are scoring test results data, weighting data scores as needed, data processing which includes reliability, superior group, distinguishing power, the difficulty level of questions, correlation of question scores, and total score of distractor quality [12]. After examining the item quality, the evaluation feasibility that has been carried out can be explained.

2 Methods

This research is a quantitative descriptive study, the research being held at Public Senior High School 1 Purwoasri. The data are in the form of scores and items for the XI MIPA 4 Japanese language mid-term exam. There are 40 multiple choice mid-term exam questions and 31 students as subjects. The research parameters include item validity, reliability, and difficulty level. Data calculations were done using the ANATES version 4 application with the following steps:

a. Open the program ANATES ver 4 by clicking twice.
b. Enter the data that has been obtained into the ANATES ver 4 program. The data entered into the program are the number of questions, students, answer choices, answer keys, and student answers.
c. Process the data; a table will appear after entering the number of students, questions, and answers. The table contains students' names, answer keys, and student answers. After all, tables are filled in, press 'process all automatically' on the scoring table.
d. The data processing results (reliability, validity, and difficulty level of the items) will come out.

2.1 Item Validity

The correlation between the total score and the items is needed to test the item validity. Figuring the validity can be done with the following formula [13]:

$$r_{xy} = \frac{N\sum XY - \left(\sum X\right)\left(\sum Y\right)}{\sqrt{\{N\sum X^2 - \left(\sum X\right)^2\}\{N\sum Y^2 - \left(\sum Y\right)^2\}}}$$

Description:
$r_{xy} r_{xy}$ = Correlation coefficient between variables X and Y
X = Item scoreY = Total scoreN = Total students
To interpret the magnitude of the correlation index, the validity coefficient is considered valid if it ranges from 0.3 to 0.5 [14]. The accuracy of the test in making measurements is the essence of validity, so the higher the coefficient value, the more accurate the test.

2.2 Reliability

A reliable test is a test that is consistent and produces a score that does not change. According to Siregar [15], the instrument can be interpreted as reliable if the reliability coefficient is more than 0.60. Below is the steps used to determine the reliability value.

$$r_{11} = \left[\frac{k}{k-1}\right]\left[1 - \frac{\sum \sigma_b^2}{\sigma_t^2}\right]$$

Description:
r_{11}= Reliability
k = Number of questions
$\sum \sigma_b^2$ = Number of item variants
σ_t^2 = Total variance

2.3 Level of Difficulty

The proportion between students' correct answers and the total number of students is a calculation to determine the difficulty level. The index used to indicate the difficulty or ease of the item uses a number between 0.00 to 1.00. Below is the formula used to determine the difficulty index of multiple-choice questions.

$$P\frac{B}{JS}$$

Description:
P = Difficulty index
B = Number of students who answered the question correctly
JS = Total number of students who took the test [16]
To interpret which level the question is, the following classification table is used:

Table 1. Classification of difficulty levels [17]

Difficulty Index	Difficulty Category
0 – 0,30	Difficult
0,31 – 0,70	Medium
0,71 – 1,00	Easy

3 Results and Discussion

The analysis is divided into three parts: validity, reliability, and difficulty level of the questions. The results revealed the test standard in the XI MIPA 4 of Public Senior High School 1 Purwoasri (Table 1).

3.1 Item Validity

The validity test aims to determine the number of valid and invalid questions. The description carried out using the application is described in the following table:

Whether or not a question is valid is determined by the significance of the correlation coefficient as a benchmark. Table 2 shows the results obtained from the mid-term exam items' validity on XI MIPA 4 students. The significance limit for the correlation coefficient from the analysis used was 0.304 at a significance level of 0.05 and 0.393 at a significance level of 0.01. 40 in Table 3 column Df (N-2) is the number of items used in the problem. The results show that there are only 2 valid items, while the remaining 38 items have invalid validity. Of the 40 questions, 18 items have a minus correlation value or can be interpreted as a minus *rxy* or *r* table (the correlation value in Table 2) is smaller than *t* count (significance limit), so the question is automatically invalid.

There are more invalid correlation values in this study, so it can be interpreted that the test does not carry out its measuring function. The test also produces data that is not in line with the measurement objectives. As a follow-up to the study results, valid questions can be reused as the next question. Meanwhile, invalid questions must be corrected with appropriate questions on the achievement indicators. Repair of questions should also be adjusted to avoid insignificance.

3.2 Reliability

Calculation of reliability is used to measure the consistency of the measuring instrument. Of the 40 multiple-choice, a reliability index of 0.08 was obtained. The calculation uses the back technique to divide between odd and even questions. According to Siregar [15], the minimum coefficient so that the questions are reliable is 0.60. Based on the benchmark, if $r11 \geq 0.60$, then the questions tested can be interpreted as questions that have high reliability. Meanwhile, for $r11 \leq 0.60$, the questions tested have low reliability. So, the reliability is included in low reliability (index of 0.08).

Table 2. The analysis results of the item validity

Question Number	Correlation	Significance
1	0.145	
2	0.279	
3	-0.133	
4	-0.168	
5	-0.027	
6	0.333	Significant
7	0.165	
8	-0.176	
9	0.269	
10	-0.105	
11	0.204	
12	0.127	
13	-0.018	
14	0.079	
15	0.265	
16	-0.018	
17	0.020	
18	0.099	
19	-0.160	
20	0.291	
21	0.280	
22	0.099	
23	-0.290	
24	-0.242	
25	0.280	
26	-0.054	
27	-0.230	
28	0.280	
29	-0.018	
30	-0.105	
31	0.097	
32	-0.018	

(*continued*)

Table 2. (*continued*)

Question Number	Correlation	Significance
33	0.145	
34	-0.290	
35	-0.128	
36	-0.176	
37	0.270	
38	-0.168	
39	0.069	
40	0.352	Significant

Table 3. Correlation coefficient significance limit

Df (N-2)	p = 0,05	P = 0,01
10	0,576	0,708
15	0,482	0,606
20	0,423	0,549
25	0,381	0,496
30	0,349	0,449
40	0,304	0,393
50	0,273	0,354

3.3 Level of Difficulty

The difficulty level of questions can be calculated by looking at the students' results. If only a few students can answer the questions, the higher the difficulty of the questions and vice versa. The summary is shown below.

The Table 3 is interpreted the difficulty level based on Sudjana's classification of difficulty levels [17]. The number that shows the difficulty level is called the difficulty index, which ranges from 0.00 to 1.00 (0% to 100%). There are 3 classifications of the difficulty index: easy (71% to 100%), medium (31% to 70%), and difficult (0% to 30%). The data obtained from Table 4 are 32 easy questions, 6 medium questions, and 2 difficult questions. The category of difficult questions is when the number of students who answer the questions correctly is less than 30% of the entire class. So, the percentage here means the percentage of students who can answer the question correctly. Below is the text of the difficult questions.

Table 4. Difficulty level analysis results

No	Correct Amount	Difficulty Level %	Interpretation
1	27	87.10	Easy
2	26	83.87	Easy
3	25	80.65	Easy
4	30	96.77	Easy
5	25	80.65	Easy
6	22	70.97	Medium
7	26	83.87	Easy
8	26	83.87	Easy
9	29	93.55	Easy
10	27	87.10	Easy
11	24	77.42	Easy
12	23	74.19	Easy
13	28	90.32	Easy
14	25	80.65	Easy
15	28	90.32	Easy
16	28	90.32	Easy
17	27	87.10	Easy
18	29	93.55	Easy
19	28	90.32	Easy
20	25	80.65	Easy
21	20	64.52	Medium
22	29	93.55	Easy
23	26	83.87	Easy
24	29	93.55	Easy
25	20	64.52	Medium
26	16	51.61	Medium
27	27	87.10	Easy
28	20	64.52	Medium
29	28	90.32	Easy
30	27	87.10	Easy
31	7	22.58	Difficult
32	28	90.32	Easy

(*continued*)

Table 4. (*continued*)

No	Correct Amount	Difficulty Level %	Interpretation
33	27	87.10	Easy
34	26	83.87	Easy
35	22	70.97	Medium
36	26	83.87	Easy
37	27	87.10	Easy
38	30	96.77	Easy
39	30	96.77	Easy
40	8	25.81	Difficult

31. Salsabila san surabaya kara Malang ..... basu de ikimasu.

a. Mimasu *d. kakimasu*
b. Wa *e. made*
c. Wo

40. Suiyoubi ni gakkou no..... wo kimasu

a. Iro *d. Taisougi*
b. Nani *e. Kinyoubi*
c. Kimasu

Both questions are difficult. The difficulty level was obtained from the number of students' correct answers; only 7–8 people answered both questions correctly. Suppose the item is included in the difficult level. In that case, there will be hypotheses, such as 1) there is an error in the answer key, 2) the item has more than one correct answer, 3) the material used is not suitable to use in multiple-choice, 4) the sentence confusing and too long, and 5) the material in question has not been completely taught. For other questions, 32 items (80%) are fairly easy because more than 70% of the children answered correctly. For medium questions, there are only 6 items (15%) of the total number of questions.

After obtaining the analysis results, the next step is to improve the items. According to Arikunto [11], a good question is a question of moderate difficulty. Easy questions will not improve students' abilities. Students are not stimulated to accept challenges since they are used to answering easy questions and vice versa.

The next step is to fix, discard or continue to use the existing questions. According to Suherman & Yaya [18], question items with a difficulty index between 21% and 80% can be used directly. Problems with a difficulty index between 81% to 90% can be used with improvement first. Then, questions with a difficulty index of 91%-100% should be replaced.

In this study, there were 5 questions with a difficulty index between 91%-100%, so these questions had to be replaced. At the index of 81%-90%, 19 items can be used as

long as they are re-examined and then corrected first. The 19 items were included in the easy level since a total of 31 students, there were more than 25 people who managed to answer the questions correctly, so they were categorized as too easy items. At the index of 21%-80%, 16 questions can be directly reused.

4 Conclusion

Based on the analysis results, the conclusions are:

a. Regarding the validity, only 2 out of 40 items were valid. The significance level limit is determined by the correlation coefficient limit of 0.304 at a significance level of 0.05 and 0.393 at a significance level of 0.01. The other 38 items have a correlation coefficient below 0.304; even 18 of them have a minus value (low significance level).
b. In reliability, the mid-term exam questions have a reliability of 0.08. This value is included in a very low category since the reliability coefficient is far below the reliable benchmark of 0.60.
c. From the difficulty level, there are 32 easy questions, 6 medium questions, and 2 difficult questions. From the percentage, 16 items were obtained that could be reused immediately, 19 items that could be reused as long as they were repaired first, and 5 items that had to be replaced since they were too easy.

For future researchers, it is recommended to examine the item analysis related to distracting and discriminating questions in more detail. The number of values that are not significant and the unreliability of the questions make researchers have to pay more attention in making questions. The improvement of these questions will impact student grades in the future.

Authors' Contributions. Rizki A. N. Dayanto was the person responsible for conceiving of and making the plans for the study. Syamsul Sodiq was the one who was responsible for carrying out the research. Miftachul Amri was helpful in determining how the data should be interpreted. Every contributor gave critical feedback that was both intelligent and constructive, and they all had a hand in both the research and the writing of the study.

References

1. M. N. Purwanto, *Prinsip-Prinsip dan Teknik Evaluasi Pengajaran.* Bandung: Remaja Rosdakarya, 2011.
2. Z. Arifin, *Evaluasi Pembelajaran.* Bandung: Remaja Rosdakarya, 2012.
3. M. Mulyadi, "Penelitian kuantitatif dan kualitatif serta pemikiran dasar menggabungkannya," *Jurnal Studi Komunikasi dan Media*, vol. 15, no. 1, pp. 127-138, 2011.
4. D. Febyronita dan G. Giyanto, "Survei tingkat kemampuan siswa dalam mengerjakan tes berbentuk jawaban singkat (short answer) pada mata pelajaran IPS terpadu (geografi) kelas VII di SMP Negeri 1 Mesuji tahun pelajaran 2015/2016," *Jurnal Swarnabhumi*, vol. 1, no. 1, pp. 17–23, 2016.
5. Depdiknas, *Panduan Penulisan Soal Pilihan Ganda*, Jakarta: Pusat Penilaian Pendidikan Balitbang-Depdiknas, 2007.

6. P. Rintayati, H. Lukitasari, and A. Syawaludin, "Development of two-tier multiple choice test to assess Indonesian elementary students' higher-order thinking skills," *International Journal of Instruction*, vol. 14, no. 1, pp. 555–566, 2020.
7. T. Novia, A. Wardani, C. Canda, N. Nurdi, dan N. Nurmasyitah, "Analisis validitas dan reliabilitas butir soal UTS fisika kelas X SMA Swasta Muhammadiyah 4 Langsa," *GRAVITASI: Jurnal Pendidikan Fisika dan Sains*, vol.3, no. 1, pp. 19–22, 2020.
8. S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan.* Jakarta: Bumi Aksara, 2013.
9. Z. Arifin, *Evaluasi Pembelajaran.* Bandung: Remaja Rosdakarya, 2013.
10. D. Sugiyono, *Metode Penelitian Kuantitatif, Kualitatif dan R&D.* Bandung: Alfabeta, 2018.
11. S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan (Revisi Cet. 6).* Jakarta: Bumi Aksara, 2006.
12. T. Harianto, *Analisis Hasil Evaluasi Pembelajaran Meliputi Daya Beda, Tingkat Kesulitan, Reabilitas, dan Keberfungsian Distraktoran dengan Software Anates.* Malang: Universitas Negeri Malang, 2014.
13. S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan (Edisi Revisi).* Jakarta: Bumi Aksara, 2005.
14. S. Azwar, *Reliabilitas dan Validitas.* Yogyakarta: Pustaka Pelajar, 2008.
15. S. Siregar, *Metode Penelitian Kuantitatif: Dilengkapi dengan Perbandingan Perhitungan Manual dan SPSS.* Jakarta: Kencana, 2017.
16. S. Arikunto, *Dasar-Dasar Evaluasi Pendidikan: Edisi 2.* Jakarta: Bumi Aksara, 2012.
17. N. Sudjana, *Penilaian Hasil Proses Belajar Mengajar*. Bandung: Remaja Rosdakarya, 2014.
18. E. Suherman dan Y. Sukjaya, *Petunjuk Praktis untuk Melaksanakan Evaluasi Pendidikan Matematika.* Bandung: Wijayakusumah, 1990.