Indonesian Cognitive Test for Educational Psychology

M. Arvani Zakky Al Kamil(✉), Iqbal Ali Wafa, and Muchamad Adam Basori

Maulana Malik Ibrahim Islamic State University, Malang, Indonesia
mbasori09@bsi.uin-malang.ac.id

Abstract. Learning evaluations need to be carried out by teachers or lecturers as educators to measure the achievement of the goals that have been set. This research was conducted as a learning evaluation in Educational Psychology lectures. This research aims to examine the differentiating power, level of difficulty, and quality of items in the Educational Psychology course at UIN Maulana Malik Ibrahim Malang. The questions in the assessment are arranged based on four themes in this material and developed into 24 items. This research was conducted at the Faculty of Psychology UIN Maulana Malik Ibrahim Malang with 19 students participating. The research was conducted by interpreting descriptive quantitative on the results of existing qualitative answers. Analysis of learning evaluation questions using Anates software to test, among others, the reliability of the questions, discriminating power, degrees of difficulty, a correlation between item scores on the total score, and the quality of distractors. The analysis results found that the learning evaluation questions could be used again. Suggestions for the research are some changes to items with low discriminatory power and an increase in the quality of distractors with higher quality items. The results of these questions have a good level of difficulty, evenly distributed in complexity, ranging from very difficult to very easy.

Keywords: Learning Evaluation · Education Psychology · Anates

1 Introduction

Evaluating student learning outcomes is one of the responsibilities of an educator. Assessment is an effort or action to determine the achievement of learning objectives that have been set. Assessment (assessment) is used to assess the success of the process and student learning outcomes. This assessment can be carried out in three domains: the cognitive, emotional, and psychomotor domains. Cognitive domain assessment can be done through several techniques. Daily exams, mid-semester tests, end-of-semester tests, and national final exams are some tests to measure the cognitive domain (Gunawan & Palupi, 2016). The exam can be done using essay questions or multiple choice. Educator evaluation is required in some circumstances to make multiple choice questions correctly and accurately and to evaluate whether the exam questions can be used in the assessment.

A. Fattah et al. (Eds.): PFH 2022, ASSEHR 728, pp. 77–89, 2023.
https://doi.org/10.2991/978-2-38476-032-9_9

Multiple choice questions must be answered by choosing one from the list multiple choice (Alwi, 2015). Multiple choice is a popular type of objective test and is often used in learning evaluation activities. Multiple-choice questions have two parts: the subject matter (stems), which consists of the problem being evaluated, and several choices or alternative responses (options). Among the many alternative answers, there is only one correct answer, called the answer key. Options other than the answer key serve as a distraction. There are various rules in the design of multiple-choice questions, namely topic, structure, and language. Only multiple choice exam questions assess the level of cognition (from memory to evaluation); the form of assessment is simple, fast, and objective, covering various levels of educational content, and easy to apply mass to many participants (Slamet & Maarif, 2014). The disadvantages of the multiple-choice test are: that it takes a long time to make questions, it is not easy to provide a homogeneous and functional distractor, and students can guess the right answer. Educators often believe that the exam questions they have prepared are good, so they hope the results obtained by students will be good too. However, in practice, it is only sometimes in line with the expectations of educators, so you must review the test results to assess the quality of the test kits and the effectiveness of the items in the test kits test.

Analysis of items in an exam needs to be done to evaluate each item to be of quality. High-quality questions can produce the right information as intended. This suggests that item analysis provides information about exam questions' quality or suggestions on how to improve their quality. According to Linn and Gronlund, the item analysis aims to solve the following questions; 1) Does the difficulty level match the question? 2) Is there anything else in the irrelevant question? 3) Do the answer choices work well? (Sidin & Khaeruddin, 2012). According to Arikunto (2010), The purpose of test analysis is to assist teachers in detecting problematic items, obtain information that can be used to improve the quality of questions to be used again in the future and obtain a brief overview of the status of the questions that have been asked. Collected. Another purpose of item analysis, according to Sidin & Khaeruddin (2012), is to categorize questions (good, bad, and in need of improvement), increase the effectiveness of alternative responses to questions (especially distractors), raise questions which need improvement, and choose a question. A good question can be used as a final preparation for an exam.

Item evaluation is categorized as qualitative and quantitative (Mansyur & Harun, 2015). Qualitative evaluation is associated with the content and form of questions. In contrast, quantitative evaluation includes evaluating the internal characteristics of the questions through empirical statistics (Mansyur & Harun, 2015) and related to the use of statistical properties by educators (Sidin & Khaeruddin, 2012). The components analyzed closely relate to the material/content material, construction, and language. Meanwhile, quantitative evaluation is mainly based entirely on empirical statistics of these items. In this method, the questions have been prepared and tested on students to achieve empirical statistics.

Therefore, it is necessary to develop learning evaluation tools as a research topic that needs special attention. Educational psychology is a subject that needs to be developed device. This research article looks at the differentiating power, difficulty level, and quality of items in the Educational Psychology course at UIN Maulana Malik Ibrahim Malang.

Table 1. .

No	Main Theme Discussion	
1.	Learning Planning	a. Definition b. Learning planning techniques
2.	Learning Implementation	a. Definition b. Learning implementation process
3.	Learning Technology	a. Definition b. Application of technology in learning
4.	Class Management	a. Definition b. Class management goals

2 Method

This study uses quantitative research methods by providing a descriptive interpretation of the existing qualitative data. Qualitative research data is processed through percentages, described, and evaluated qualitatively. The results of quantitative data are processed using ANATES software to describe the results of student responses as research findings (Arif, 2014). The analysis carried out includes the reliability of the questions, the differentiating power of the questions, the level of difficulty, the correlation of scores items on the total score, and the quality of the distractor questions. The research was conducted at the Faculty of Psychology, UIN Maulana Malik Ibrahim. This research was conducted as an evaluation of the Educational Psychology course. The number of participants in this study was 19 students.

The number of questions arranged in this device is 24 questions. The questions developed are divided into four themes, with a total of two subjects for each theme. The number of subjects is 21 subjects. The question grid is shown in Table 1.

3 Results

3.1 Reliability Question

The purpose of this study is to see the level of trustworthiness of the questions that have been structured (reliable). When given to the same group many times and on different occasions, a test is considered reliable if the findings are always the same. If a question learning evaluation is unreliable, who should not give it again at the next evaluation.

The results of consistency or stability of the measurement results of exam questions are called test reliability (Bhakti, 2015). Reliable measuring tools can produce consistent scores when used to test the same object repeatedly. The reliability coefficient or standard measurement error is a parameter based on the reliability coefficient. As a measure in general to represent the reliability of the test. The halving technique, which uses the Spearman-Brown formula to calculate the reliability of all tests (Eisinga et al., 2013), is one way to find the value of the reliability coefficient. The results of the question reliability test are shown in Table 2.

Table 2. Question Reliability

Content	Value
Average	15.79
intersection Baku	1.99
Correlation XY	0.09
Test Reliability	0.17

The maximum score for this assessment is 24. The number of questions in this assessment is 24 items questions. The number of participants in this evaluation is 19 students. The analysis of student answers shows that the average score obtained by students is 15.79, with a standard deviation of 1.99. The reliability test results showed that the test reliability score was 0.17. If the value is greater than or equal to 0.17, it indicates that the question of learning outcomes being tested can be concluded if it has high reliability. If the value is less than 0.17, it indicates that the question of learning outcomes being tested can be concluded if it has low reliability.

3.2 Power Different

The discriminatory power of the questions is the test's ability to distinguish between students who are smart or have high abilities and those who are less intelligent or have low abilities (Solichin, 2017). The power of difference can be calculated using the following equation:

$$\mathrm{D} = \frac{BA}{JA} - \frac{BB}{JB}$$

Information
D: Differential power
JA: The number of participants in the upper group
JB: The number of participants in the group lower
Ba: The number of participants from the upper group chose the correct answer
Bb: The number of participants from the lower group chose the correct answer

The criteria for distinguishing power (DB) are shown in Table 3. While the results of the discriminatory power test are shown in Table 4.

The results of the above test are 1 question with very good discriminating power, six questions with good discriminating power, seven questions with sufficient discriminating power, nine questions with poor discriminatory power, and 1 question with poor discriminating power (must be discarded). From the discriminatory power test results, it can be seen that questions number 1 and 4 are questions that have bad distinguishing power, so they must be discarded.

3.3 Level Difficulty

Analysis of the level of difficulty of the question aims to assess an item of the question, including easy or difficult criteria (Boopathiraj & Chellamani, 2013). A difficulty level

Table 3. Criteria for distinguishing power (DB)

Power criteria Different DP	Qualification Questions
<0.00	Very bad, must thrown away
0.00 s/d 0.19	Bad
0.20 s/d 0.39	Enough
0.40 s/d 0.69	Fine
0.70 s/d 1.00	Very Well

is a number that indicates the ease or difficulty of a question item (Arikunto, 2010). The result of a number that shows how difficult or easy something can be is called the level of difficulty (Arikunto, 2003). Level, The difficulty of the items, is calculated through the equation:

$$P = \frac{\mathrm{B}}{\mathrm{J}}$$

Information:
P: difficulty index,
B: number of examinees who chose the correct answer, and
J: number of participants exam.

The results of the calculation of the degree of difficulty of the questions that have been made are shown in Table 5.

From the results of the level of difficulty test, it can be seen that there are three questions in the very difficult difficulty category, one question in the difficult difficulty category, seven questions in the moderate difficulty category, five questions in the easy difficulty category, and eight questions with very easy difficulty category. This shows that the questions' difficulty level is relatively evenly distributed from the very difficult to very easy categories.

3.4 Item Score Correlation with Score Total

A test item is valid if it strongly supports the overall score (Alpusari, 2014). Item test items are valid if they greatly support the total score (Masriyah, 1999). The overall score may be high or low depending on the results of each test, and the validity of a test item is said to be high if the item score is correlated with the total score. This accuracy can be evaluated using the correlation formula to determine the validity of the test items. Objective form questions generally have a score of 1 (for correct answers) or 0 (for incorrect answers), with the overall score calculated by adding up the scores for each item in the test set. The product-moment correlation formula is used to calculate the magnitude of the correlation coefficient, namely:

Table 4. Results of the Differentiating Power Test

No Item	Top Group	Bottom Group	Different	DB Index (%)	Qualification
1	3	3	0	0.00	Bad
2	5	4	1	20.00	Enough
3	5	4	1	20.00	Enough
4	1	1	0	0.00	Bad
5	4	3	1	20.00	Enough
6	5	5	0	0.00	Bad
7	0	0	0	0.00	Bad
8	5	4	1	20.00	Enough
9	5	2	3	60.00	Well
10	4	2	2	40.00	Well
11	5	4	1	20.00	Enough
12	5	2	3	60.00	Well
13	2	0	2	40.00	Well
14	0	2	−2	−40.00	Must Throw
15	4	2	2	40.00	Well
16	5	5	0	0.00	Bad
17	0	0	0	0.00	Bad
18	5	2	3	60.00	Well
19	4	4	0	0.00	Bad
20	5	5	0	0.00	Bad
21	3	2	1	20.00	Enough
22	5	4	1	20.00	Enough
23	5	5	0	0.00	Bad
24	5	1	4	80.00	Very good

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

where:

r_{xy}: product moment correlation coefficient
X: score items
Y: overall score
N: the number of students who take the exam

Table 5. Result of Question Difficulty Level

No Item	Correct Amount	Difficulty Level (%)	Information
1	11	57.89	Currently
2	18	94.74	Very easy
3	14	73.68	Easy
4	8	42.11	Currently
5	12	63.16	Currently
6	18	94.74	Very easy
7	1	5.26	Very Difficult
8	18	94.74	Very easy
9	15	78.95	Easy
10	14	73.68	Easy
11	18	94.74	Very easy
12	10	52.63	Currently
13	4	21.05	Hard
14	2	10.53	Very Difficult
15	13	68.42	Currently
16	19	100.00	Very easy
17	0	0.00	Very Difficult
18	16	84.21	Easy
19	16	84.21	Easy
20	19	100.00	Very easy
21	12	63.16	Currently
22	18	94.74	Very easy
23	17	89.47	Very easy
24	7	36.84	Currently

From the results of the correlation analysis of the item scores with the total score, 2 questions have a very significant correlation, and 9 questions have a significant correlation (Table 6).

Table 6. Item Score Correlation Results with Total Score

No Item	Correlation	Significance
1	0.072	-
2	0.340	-
3	0.182	-
4	0.038	-
5	0.199	-
6	−0.026	-
7	0.269	-
8	0.340	-
9	0.478	Significant
10	0.367	-
11	0.340	-
12	0.550	Very Significant
13	0.390	Significant
14	−0.583	-
15	0.336	-
16	-	-
17	-	-
18	0.550	-
19	0.177	-
20	-	-
21	0.255	-
22	0.462	Significant
23	−0.303	-
24	0.534	Very Significant

3.5 Quality Distractor

Calculation of the number of examinees choosing each question answer option can be used to determine the success of each question choice. In addition, which distractors are successful, which are less or less effective, and which are deceptive can be noticed. If the majority of examinees choose a particular distractor answer while only a few choose the key answer, maybe the teacher is wrong in determining the answer key and the distractor is really the answer key. However, it is possible that the answer key is correct, but the distracting option is very interesting for missed (Table 7).

Table 7. Distractor Quality Test Results

No Item	a	b	c	d	*
1	4+	4+	11**	0--	0
2	0--	1---	0--	18**	0
3	5---	0--	14**	0--	0
4	2+	0--	9---	8**	0
5	18**	12**	0--	0--	0
6	7---	---	0--	0--	0
7	3-	11--	1**	4+	0
8	1---	0--	18**	0--	0
9	0--	15**	1+	3---	0
10	0--	5---	0--	14**	0
11	0--	18**	0--	1---	0
12	10**	9---	0--	0--	0
13	9--	3+	3+	4**	0
14	1--	15---	2**	1--	0
15	13**	1-	3+	2++	0
16	0	0	0	19**	0
17	0--	0**	5++	14---	0
18	16**	0--	3---	0--	0
19	1++	16**	0--	2--	0
20	19**	0	0	0	0
21	1-	1-	12**	5---	0
22	0--	18**	0--	1---	0
23	0--	17**	1+	1+	0
24	7**	1--	0--	11---	0

Information:
**: Correct answer
++: Very Good
+: good
-: Not good
--: Bad
---: Very bad

Based on the data above, some of the distractors are still categorized as bad and very bad. This distracting answer choice needs to be further improved to improve the quality of the learning evaluation questions.

4 Discussion

In the evaluation of learning, making a test or how to measure it, its implementation, and how to interpret it cannot be ignored. A test is a measurement tool that provides information about students. There are several kinds of tests, and based on these tests, educators obtain information about their students, which is the basis for making decisions that can determine the fate of these students. The educational process is one way of transferring knowledge in a structured manner. In short, providing knowledge is better known as the learning process. This process is not just transferring knowledge, there are several evaluations to determine how much students receive from the results of the learning process. There is a stage of results assessment or learning assessment in the learning evaluation process. Cognitive assessment is an assessment that is often used. This is because it relates to students' ability in the learning process in educational institutions. The cognitive aspect is related to the mastery of knowledge. The affective aspect is related to scientific values and attitudes.

In measuring the cognition of students, a kind of achievement test is needed as a form of learning evaluation. So, several questions were compiled as a form of application of process assessment and learning evaluation devoted to measuring students' cognitive domain in educational psychology courses. Preparing questions or tests should include the following aspects: test planning, test implementation, and results management. To determine which learning outcomes or levels of thinking ability will be assessed, the test compiler can be guided by the instructional objectives or the evaluation objectives themselves. Before compiling the questions, the test writer needs to make a test grid as a very important thing. Test grids can provide valid and reliable information.

Arikunto (2014) explains the need to analyze the questions asked to determine which questions are very good, somewhat bad, and bad so that questions that are considered rather bad or bad can be corrected. Judging from the analysis of the difficulty level, items included in the good category (in the sense that the degree of difficulty of the items is sufficient or moderate) must be immediately stored in the question book. Then the questions can be used again to evaluate future learning outcomes. Examiners must re-examine and evaluate questions in easy categories to know the elements that make the questions able to be answered by almost all test takers (Asri & Burhan, 2014).

Suppose the question item belongs to the easy category. In that case, it can be predicted that the item is: a distracting question that does not work or that most students correctly answer the item, which means that most students already understand the material being asked. As for questions included in the difficult category, the examiner must re-examine, explore, and evaluate the things that make the questions difficult (Kadir, 2015). If a question item is in a difficult category, the following are likely to happen: the item may have an incorrect answer key; the item has two or more correct answers; the question item in question has not been submitted, or the learning has not been completed, the factor is that the minimum competence of students has not been achieved; the measured ability is not by the form of the question being tested; or the question sentence is too long and complex (Kadir, 2015).

From the results of the test reliability analysis using the halving technique, which uses the Spearman-Brown formula (Eisinga et al., 2013), it can be seen that the test reliability score that has been made is 0.17. If this value is more than or equal to 0.70,

then it can be seen that what has been made is tested to see if it has high reliability. A high-reliability score indicates that the measurement through questions made is consistent or stable (Bhakti, 2015). Reliable questions will produce consistent results when repeatedly testing the same material. The reliability coefficient or standard measurement error is a parameter derived from the reliability coefficient and is a common way to represent the reliability of the test. From the correlation analysis of item scores on the total score, two questions have a very significant relationship, and seven questions have a significant correlation. Some questions have good significance so that they can be used repeatedly as material test questions in future evaluations.

The differentiating power obtained from the results of Anates' analysis has a level of very good, good, sufficient, bad, and not good. Based on the test results above, there is 1 question with very good discriminating power, six questions with good discriminating power, seven questions with sufficient discriminating power, nine questions with poor discriminating power, and 1 question with poor discriminating power (must be discarded). From the discriminating power test results, it can be seen that question number 1 4 is a question that has poor distinguishing power, so it must be thrown away. The power of difference makes the test items able to classify students according to their level of cognitive ability to understand educational psychology material.

From the test results of the level of difficulty of the questions, it can be shown that there are three questions with a very difficult difficulty category, 1 question with a difficult difficulty category, seven questions with a moderate difficulty category, five questions with easy difficulty category, and eight questions with very easy difficulty category. This shows that the questions' difficulty level is relatively evenly distributed from the very difficult to very easy categories. The difficulty level can be a measure of the success of educators in seeing the competencies that students in learning educational psychology courses have mastered. The average questions have an easy difficulty category, meaning that students who take educational psychology courses can master the material and understand it well so they can work on test questions.

Based on the data above, some distractors are still categorized as bad and very bad. This distracting answer choice requires improvement so that it can improve the quality of the questions which has been made. The distractor's answer, which has been able to carry out the expected function properly, can be used again in future assessments. In contrast, the distractor's answer, which has yet to be able to carry out its function properly, can be revised or replaced with a distractor's answer. Other. The distractor's answer makes the critical thinking power of students increasingly honed to see which correct answer is meant and is by the expectations of learning in educational psychology courses as well as optimal learning evaluation.

5 Conclusion

Based on the analysis and explanation above, it is concluded that the test items that have been made can be used again. Some improvements are needed, such as replacing questions with bad discriminating power. The quality of the distractors in the answer options needs to be improved so that the quality of the questions will be of higher quality. As for the difficulty, the level matter is good. The difficulty level of the questions has

been evenly distributed from very difficult to very easy. Questions with good quality regarding validity, reliability, difficulty level, power difference, and distractors can then be collected into a set of questions for the next learning evaluation tests.

Educators should develop evaluation tools and analyze these tools because it requires measurement of students' understanding of certain lecture materials. Measurement will produce accuracy if the quality evaluation tool is well and efficiently implemented. Using proper measurement tools, the evaluation results can describe student understanding so that educators can understand the achievement or successful learning that has been carried out.

References

Alpusari, M. (2014). Analisis butir soal konsep dasar IPA 1 melalui penggunaan program komputer anates versi 4.0 for Windows. Primary: Jurnal Pendidikan Guru Sekolah Dasar, 3(2), 106–115.

Alwi, I. (2015). Pengaruh Jumlah Alternatif Jawaban Tes Obyektif Bentuk Pilihan Ganda Terhadap Reliabilitas, Tingkat Kesukaran Dan Daya Pembeda. Faktor Exacta, 3(2), 184–193.

Arif, M. (2014). Penerapan aplikasi anates bentuk soal pilihan ganda. Edutic-Scientific Journal of Informatics Education, 1(1).

Arikunto, Suharsimi. (2003). Prosedur Penelitian, Suatu Praktek. Jakarta: Bina. Aksara.

Arikunto, S. (2010). Metode peneltian. Jakarta: Rineka Cipta.

Arikunto, S. (2014). Dasar-Dasar Evaluasi pendidikan: Prosedur Penelitian Suatu Pendekatan Praktik. Jakarta: PT Bumi Aksara.

Asri, A. F., & Burhan, A. (2014). Analisis Tingkat Kesukaran, Daya Pembeda dan Fungsi Distraktor Soal Ujian Semester Ganjil Mata Pelajaran Produktif di SMK Negeri 1 Indralaya Utara Tahun Pelajaran 2012/2013. Jurnal Pendidikan Teknik Mesin, 1(2).

Bhakti, Y. B. (2015). Pengaruh jumlah alternatif jawaban dan teknik penskoran terhadap reliabilitas tes. Formatif: Jurnal Ilmiah Pendidikan MIPA, 5(1).

Boopathiraj, C., & Chellamani, K. (2013). Analysis of test items on difficulty level and discrimination index in the test for research in education. International Journal of Social Science & Interdisciplinary Research, 2(2), 189–193.

Eisinga, R., Te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? International Journal of Public Health, 58(4), 637–642.

Gunawan, I., & Palupi, A. R. (2016). Taksonomi Bloom–revisi ranah kognitif: kerangka landasan untuk pembelajaran, pengajaran, dan penilaian. Premiere Educandum: Jurnal Pendidikan Dasar Dan Pembelajaran, 2(02).

Kadir, A. (2015). Menyusun dan menganalisis tes hasil belajar. Al-TA'DIB: Jurnal Kajian Ilmu Kependidikan, 8(2), 70–81.

Mansyur, S., & Harun, R. (2015). Asesmen pembelajaran di sekolah: Panduan bagi guru dan calon guru. Yogyakarta: Pustaka Pelajar.

Masriyah. 1999. Validitas dan Realibilitas. Surabaya : Unesa University Press

Sidin, A., & Khaeruddin, K. (2012). Evaluasi Pembelajaran. Universitas Negeri Makassar.

Slamet, S., & Maarif, S. (2014). Pengaruh bentuk tes formatif assosiasi pilihan ganda dengan reward dan punishmentscore pada pembelajaran matematika siswa SMA. Infinity Journal, 3(1), 59–80.

Solichin, M. (2017). Analisis daya beda soal, taraf kesukaran, validitas butir tes, interpretasi hasil tes dan validitas ramalan dalam evaluasi pendidikan. Dirasat: Jurnal