International Journal of Innovation, Creativity and Change. www.ijicc.net
Volume 8, Issue 1, Special Edition, ICOVET, 2019

Measure critical thinking ability: validity and reliability multiple choice test

S I Haryudo[1,2], Ekohariadi[2], Munoto[2], L Nurlaela[2], B I G P Asto[2], I Wayan Susila[2]

[1]Electrical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, East Java, Indonesia
[2]Vocational Education, Postgraduate Program, Universitas Negeri Surabaya, East Java, Indonesia

Corresponding Author: subuhisnur@unesa.ac.id, ekohariadi@unesa.ac.id, munoto@unesa.ac.id

Critical thinking skills are one of the skills that must be possessed by students as graduate competency standards. Instrument evaluation is needed to evaluate critical thinking skills. This study aims to examine the quality characteristics of multiple choice test instrument items to measure critical thinking skills. The test instrument developed was an Electrical Installation Automation test instrument in order to maintain the quality of critical thinking test instruments with the addition of a wider variety of reasoning aspects. The test instrument is in the form of multiple choices which have analysis questions (C4), synthesis (C5), and evaluation (C6). The method of developing research uses the Research and Development (R & D) method. The object of the research was all the answers to the results of the tests of 40 Electrical Engineering students. Data is collected by the method of observation, interview and documentation. Data analysis techniques are qualitative and quantitative using the Test Analysis Program (TAP) application. The results showed that: The instruments developed had good characteristics in terms of validity with the Aiken index content validity (Aiken V) of 0.820 and reliability coefficient of 0.843 in a good (reliable) range. Furthermore, for the difficulty level of 0.560 in the medium category, the average differentiation of 0.333 can be categorised as

good and the results of the analysis of the distractor quality are 39 tests that function well.

Keywords: *item instrumentation, critical thinking, TAP, automation of electrical installation*

Introduction

One reason for the low quality of education is the use of inaccurate evaluation and assessment instruments. The use of instruments according to Arifin [2] has functions and roles that are very important in order to know the effectiveness of the learning process. For that reason, correct assessment will provide the right information and encourage increasing motivation and achievement in learning [6].

Instrumentation is a tool used to measure an object of research, therefore the instrument must meet a high standard of criteria. Good instrumentation must at least have valid and reliable requirements. Besides fulfilling valid and reliable requirements it must also pay attention to the characteristics of the item, namely the level of difficulty, power difference, and functionality of distractors [3, 12]. This is in accordance with the opinion of Mansyur, Rasyid & Suratno [9] who said that in order to obtain accurate information, a valid and reliable instrument was needed.

The results of observations conducted by researchers on the Electrical Installation Automation test competency instrument for students of Electrical Engineering in Surabaya State University stated that: (1) Test instructions used are not an empirical analysis (test validity, reliability, level of difficulty, power difference, and distractors); and (2) The test instruments used need an increase in student competencies, especially in the aspect of reasoning. The increase was carried out to determine the ability to make decisions in solving problems based on considerations of a wider sample. The addition of reasoning aspects can be done by the process of evaluating and evaluating learning outcomes that can help students improve their critical and creative thinking skills. The results of the assessment will make it easier for educators to obtain information on the extent to which students learn about a subject matter.

In each learning process at any level there will always be a process of evaluation or assessment. The purpose of holding evaluations and assessments according to Sudijono [12] is: "To obtain evidence data that will be a guide to where the results of the level of ability and success rate of students in

achieving learning objectives after they take the learning process". The evaluation process is very important for the development of the quality of education in order to control the quality of education [7, 9]. The Department of Defence Education Activity (DoDEA) formulates seven skills that must be achieved in 21st Century Teaching, Learning, and Leading (21st CTLL) to prepare students for 21st century success, namely critical thinking and problem solving, collaboration across networks, agility and adaptability, initiative and entrepreneurship, effective oral and written communication, accessing and analysing information, and curiosity and imagination [1, 4, 5]. One of the skills that must be possessed by students according to the DoDEA is critical thinking skills.

Critical thinking skills are part of high-level thinking skills that can create strong thinkers and reliable problem solvers [14]. Improving high-level thinking skills is closely related to the process of evaluating learning and assessment [5, 15]. Assessment can be done on one of them with a test instrument. The test is defined as an instrument or systematic procedure for observing and explaining one or several characteristics of students using a numerical scale or classification scheme [2]. The development of a critical thinking ability test instrument has a level of analysis questions (C4), synthesis (C5), and evaluation (C6). Aspects Test instruments for measuring critical thinking skills are test instruments which include reasoning for the size of concentration, reasoning for uncertainty, reasoning towards the sample, and reasoning for the association. Development of test instruments can use the form of essay tests (description) and multiple-choice tests. In this study, the form of multiple-choice test instruments was developed. This is as expressed by Betsy Moore & Tood Stanley [11] that "analysis question, synthesis, question, and evaluation of the question can also be asked using a multiple-choice format". This means that questions in the form of analysis, synthesis and evaluation can be used with multiple choice questions.

Based on the above explanation, the problem of this study is how is the quality of the multiple-choice test instrument items to measure critical thinking skills on the competency test material for Electrical Installation Automation? To measure the quality of items of multiple-choice test instruments, test items were analysed in general [8, 16]. The analysis is intended to determine the size of the index of the level of difficulty, the index of different power and the effectiveness of the deception of the items in question. One of the test quality analysis can use the Test Analysis Program (TAP). The TAP application is an instrument test program that can be used in measurements to analyse the quality of

an instrument's grain. The results of the analysis can be used as a source of accurate information and as a basis for decision making, whether the instrument is good or not.

Methods

The method used in this study is the method of research and development [8]. In this study, the form of the test instrument to be developed is an objective test in the form of a multiple-choice high-order tests (Higher Order Thinking Skills) with five answer choices. The research subjects in the development stage of this test instrument were 40 Electrical Engineering students in 2015. The objects of the research were all answers to the multiple-choice test results of the Automated Electrical Installation competency test. To obtain data in the field, researchers use several techniques, namely: (1) observation, namely data collection carried out by conducting direct observations in the field, which have to do with the problem of this research; (2) interviews, namely data collection conducted by conducting direct question and answer, which has to do with the problem of this research; (3) documentation, namely collecting data through written references in the form of answers to student test results, books, teaching materials, and others that are very relevant.

Data analysis techniques are carried out qualitatively and quantitatively. Qualitative analysis uses the expert review format and quantitative analysis by displaying the results of item response theory, namely by looking at the level of difficulty, different power and effectiveness of each item through the TAP application program. To find out the expert agreement, the validity index proposed by Aiken (1985) can be formulated as follows [12]:

$$V = \frac{\sum s}{(n\,(c-1))} \quad (1)$$

with V is an index of expert agreement regarding the validity of the contents of the item; s is the expert's score minus the lowest score in the category used ($s = r$ - I0), with r is the expert choice category score and I0 is the lowest score in the rating category); n are many experts; and c are the number of categories that can be chosen by experts. The Aiken V index value ranges from 0 to 1 with a high validity level category if $\boldsymbol{V} > 0.8$; moderate validity if $0.4 < \boldsymbol{V} \leq 0.8$; and the level of validity is less if $\boldsymbol{V} \leq 0.4$.

To produce a good test instrument, the characteristics of the other test instruments must also be known, namely differentiation, level of difficulty, deception, and reliability. To facilitate the calculation of differentiating power, give a general formula with the following formula [9]:

$$D = \frac{\sum X_A}{n_A} - \frac{\sum X_B}{n_B} \quad (2)$$

D: Index of different items, $\sum X_A$: The number of test participants answered correctly in the upper group, $\sum X_B$: Number of test participants who answered correctly in the lower group, n_A : number of test participants in the upper group, n_B: number of test participants in the lower group.

Table 1. Item Difference Index

Different Index	Categorize
$0.40 \leq D \leq 1.0$	Very good
$0.30 \leq D < 0.40$	Good
$0.20 \leq D < 0.30$	Sufficient and need a little revision
$D < 0.20$	Not good

Proportion is correct (index correct) and is the index of the difficulty of the simplest questions and is often used in determining the index size. The equation to determine the magnitude of the difficulty index is mathematically as follows:

$$\boldsymbol{P_i = \frac{\sum X_i}{Smi.N}} \quad (3)$$

Pi: the level of difficulty of the item, $\sum Xi$: Number of test takers who answer correctly, Smi: Maximum score, N: Number of test takers. The criteria used to determine the type of difficulty of the items are presented in Table 2.

Table 2. Level of difficulty of Test items

Value index	Category
P > 1.0	Easy
0.30 ≤ P ≤ 0.70	Medium
P < 0.30	Difficult

Each multiple-choice test has one question and several answer choices. Among the choices of answers, only one is correct. Besides the correct answer, there is also a wrong answer, known as a distractor. To find out the effectiveness of the trick is how well the wrong choice can trick the test participant who does not know the answer key that is available. The more test participants who choose the deceiver, then the deception can perform its function properly. A good criterion for deception is if the trickster is chosen by at least 5% of the test participants [8, 10, 12]. Furthermore, the criteria used in determining reliability using the alpha cronbach technique approach with the reliability coefficient: Alpha <0.7: inadequate, Alpha> 0.7: good, and Alpha> 0.8: excellent.

Results and Discussion

Based on the theory and the results of the empirical trial, an analysis of the test instrument was conducted aimed at measuring the quality of the instrument items in critical thinking skills in the Automation competence of Electrical Installation. Preliminary studies that have been carried out, obtained information about the concepts or theories relating to the product that will be developed such as the understanding and criteria of a test, the steps to develop the instrument, and the application of TAP in the analysis of instrument items. In addition, the most important information is obtained, namely the types of tests commonly used by educators and learning indicators.

In this study, the form of the test developed was a multiple-choice objective test. This test instrument was developed with as many as 50 items with five answer choices. The form of this test is already commonly used and has the advantage of being able to provide high objective results and can be trusted, in correcting the answers it is not difficult, and more representative in terms of covering and representing the material that has been taught. However, the multiple-choice form also has the disadvantage of being unable to measure or uncover a high or deep-thinking process, it is possible for participants to play speculations and guess answers, and making the problem is not as easy as an essay test [13].

To overcome the weaknesses of multiple choice tests, especially high-level thinking processes, variations in instrument items need to be developed. The development of instrument items in the aspect of reasoning is to provide greater mastery, so that the multiple choice forms remain of good quality. According to Azwar [3], multiple choice forms can also reveal high-level thinking skills, because this form has the most extensive variety of mastery levels, ranging from the simple to the highest.

In Bloom's Taxonomy, at the level of the C4 question (analysis) the operational verb used is analysing, auditing, solving, concluding, linking, and detecting. For the level of C5 questions (synthesis) the operational verbs used are combining, formulating, categorising, creating, designing, and compiling. Meanwhile, for the level C6 (evaluation) the operational verbs used are weighing, criticising, comparing, inferring, evaluating, predicting, and interpreting [4, 5, 14].

To determine the suitability of the designed test instruments and critical thinking skills to be measured, content validation was carried out by involving three experts. Each expert provides an assessment and suggestion for the instrument that has been designed in the content validation assessment sheet. To find out the results of the agreement of three experts, the analysis of the Aiken V index was used. The results of the expert agreement based on the Aiken V index are presented in the following table 1

Table 3. Results of the Aiken V Index Analysis on the Validation of the Content of the Critical Thinking Skills Test

No	Aspects	Indicators	Index of Aiken V	Interpretation	Average and Interpretation
1	Material	a Compliance with the questions indicator	0.83	High validity	0.83 High validity
		b Conformity with aspects of critical thinking skills 0.83	0.83	High validity	
2	Construct	a. The subject matter is formulated briefly, clearly, and firmly.	0.83	High validity	0.83 High validity
			1.00	High validity	
		b. There is only one answer key	0.83	High validity	
		c. Choice of homogeneous and logical answers	0.83	High validity	
		d. Images, graphics, tables or the like are clear and functional	0.83	High validity	
		e. The subject matter is free from multiple statements	0.67	Moderate validity	
		f. The choice of answer in the form of numbers is arranged in the order of the size of the number			
3	Language	a. Item formulation uses communicative language	0.75	High validity	0.79 Moderate validity
		b. Question items use standard Indonesian and are in accordance with EYD	0.83	High validity	
		Average			0.820 High validity

Based on table 3, it is known that the Aiken V index of the critical thinking skills test instrument is 0.820 with a high validity category. That is, the three experts stated that the instruments for testing critical thinking skills designed by researchers were worthy of testing, so that their characteristics could be known. Furthermore, with table 3 obtained KR 20 which resembles Cronbach's Alpha coefficient shows the number 0.843 which means items are close to 1.00, then overall item reliability is in a good (reliable) range. The Standard Error of Measurement (SEM) is an estimation of the standard deviation of measurement errors for each test / scale. The SEM score of KR20 in table 3 shows 2,883, the better the reliability of the test score the smaller the SEM score and the more we believe or believe in the accuracy of a test.

The results of the trial data are obtained from the answers of the corrected students by using dichotomous data assessments that have a value of 1 if the answer is correct and 0 if the answer is wrong, so that the results of the trial data in the form of numbers are obtained. The numbers are then analysed using the theory of item responses with the help of TAP 19.1.4, so that the characteristics of the test items are known. The results of the descriptive analysis for all items can be seen in Table 4.

Table 4. Results of Descriptive Analysis of Test Scores

Criteria for	Analysis of Results
Number of Examinees	40
Total Possible Score	50
Minimum Score	12 = 24.0%
Maximum Score	46 = 92.0%
Median Score	29 = 58.0%
Mean Score	28 = 56.0%
Standard Deviation	7.284
Skewness	-0.14
Kurtosis	0.019

Based on the results of the analysis in Table 5, it appears that the number of respondents who took this test were 40 people. If a test participant answers all questions correctly then the maximum score that might be obtained is 50. However, from the test results, the maximum score obtained by the

respondent is 46, the minimum score is 12 with an average score of 28, standard deviation 7.284 and variance at 53.05. The value of skewness and kurtosis which is close to zero indicates that the learning outcome test scores are normally distributed. Furthermore, the results of descriptive item analysis can be seen in Table 6.

Associated with Table 6. It appears that the number of items analysed were 50 items and 11 items there were potential items that could be eliminated because they did not meet the requirements for the item difficulty level (0.2 <p <0.95), item power index different items (D <0), pbis <0, adjp>> 0. The results of the analysis found that the average level of difficulty of items was 0.560 which indicates that the items used in the student competency test were included in the medium category, meaning that the instrument items were not too easy and not too difficult. Furthermore, the difference in grain power was calculated based on the division of two groups of test participants, namely the upper and lower groups.

Criteria	Hasil Analisis
Number of Items Analysed	50
Mean Item Difficulty	0.560
Mean Discrimination Index	0.333
Mean Point Biserial	0.339
Mean Adj. Point Biserial	0.289
KR20 (Alpha)	0.843
KR21	0.783
SEM (from KR20)	2.883
# Potential Problem Items	11

The upper group is categorised as participants who are classified as smart children while the test participants in the lower group are categorised participants who are classified as less intelligent. With the average power difference of 0.333 can be categorised as good with a different power index of 0.3 ≤ D <0.4 that the average item is able to distinguish between smart students with less intelligence. To find out the size of the difficulty, the difference between power is shown through the comparative results of the item discrimination analysis with the existing criteria in table 6.

Table 6. Results of Differential Analysis

Category	Items	Amount
Very good 0.4 ≤ D ≤1.0	1, 2, 3, 6, 7, 22, 28, 32, 33, 35,36, 37, 38, 41, 42, 43, 44, 45, 47, 48	20 (40%)
Good 0.3 ≤ D < 0.4	2, 5, 8, 11, 15, 18, 19, 26, 29, 39, 40, 46	11 (22%)
Sufficient (need a little revision) 0.2 ≤ D < 0.3	4, 7, 13, 17, 20, 21, 23, 27, 31, 49	10 (20%)
Not good D < 0.2	9, 12, 14, 16, 24, 25, 30, 34, 50	9 (18%)
	Item Total	50 (100%)

Based on the results of the analysis in table 6, it appears that the number of items that have a very good difference in power is 20 items (40%) and good as many as 11 items (22%) which means that items can be used without revision. In addition, there are also items that have different power in sufficient categories with as many as 10 items (20%) which means they have to go through the revision stage. If these items have been revised, the items can be used. While the number of items in the bad category is 9 items (18%) so that they can be considered not used. To complete the previous statement, the results of the analysis of the difficulty level of each item can be seen in Table 8.

Table 8. Results of Difficulty Level Analysis

Category	Items	Amount
Easy p > 0.7	2, 3, 6, 12, 14, 15, 16, 19, 20, 26, 28, 32, 33, 36, 37, 41, 43, 44, 48	19 (38%)
Moderate 0.3 ≤ p ≤ 0.7	1, 4, 7, 8, 9, 10, 11, 17, 21, 22, 24, 25, 29, 30, 35, 38, 42, 45, 46, 47, 49, 50	22 (44%)
Difficult p < 0.3	5, 13, 18, 23, 27, 31, 34, 39, 40	9 (18%)
	ItemTotal	50 (100%)

Referring to table 8, the analysis of instrument items in the difficult category was 9 items (18%), items in the medium category were 22 items or equal to (44%) while the items in the easy category were 19 items (38%).

From the results of the analysis it appears that the number of items in the competency test instrument is spread evenly, but it should be noted that the results of the analysis are not the only indicator that items that are difficult or easy are bad, because in the analysis of items there are several categories to consider such as difficulty and different power to assess whether the item is good or not. The results of different power analysis can be seen in table 8.

Furthermore, to see the function of the deception of the items that function in table 9, the low difference of power is usually caused by the level of function of the deception of the item, besides that the deceiver also has an impact on the level of difficulty of the item because if there are one or two deceivers on an item that is not the function, the index level of difficulty of the items will decrease, because the chances of the test participants to answer correctly are increasing.

Based on table 9 from the results of the analysis on item number 9, that the laying of the answer key located in option 1 (A) cannot function properly and it is assumed that option 4 (D) is better because the majority of the test takers consider that option 4 (D) as the answer key, different from option 1 (A), which is chosen by the test participants is less clever so it must be revised / discarded.

Table 9. Location of the Answer Key

Item	Group	Option 1	Option 2	Option 3	Option 4	Option 5
9	TOTAL	16*(0.400)	1 (0.025)	2 (0.050)	20 (0.500)	1 (0.025)
	High	4 (0.286)	1 (0.071)	0 (0.000)	9 (0.643)	0 (0.000)
	Low	6 (0.545)	0 (0.000)	2 (0.182)	3 (0.273)	0 (0.000)
	Diff.	-2 (-0.260)	1#(0.071)	-2 (-0.182)	6# (0.370)	0#(0.000)

In addition to item number 9, there are also other items that have the same thing, as shown in table 10.

Table 10. Functionality of Distractors

Option Answer Key	Number of items (counter)	Items that do not function properly	Items that function properly
1/A	11	9, 25, 31, 40	32, 33, 36, 44, 45, 47, 49
2/B	15	23, 27	2, 3, 6, 8, 11, 12, 16, 17, 19, 28, 29, 30, 43
3/C	11	4, 14	1, 15, 20, 21, 22, 26, 39, 42, 48
4/D	11	13, 18, 24	5, 7, 10, 35, 37, 38, 46, 50
5/E	2	-	34, 41

As shown in table 10, the number of items that did not function properly were as many as 11 items because of the deception which was supposed to be used as a tool to trick the test participants and was considered to be the answer key because the deceivers were chosen by the test participants. Overall suggestions are given for laying options as shown in table 11.

Table 11. Option Selection as the Answer Key Location

Item	Key Layout Item *	Substitute Advice #
9	Option 1(A)	Option 4(D)
25	Option 1(A)	Option 5(E)
31	Option 1(A)	Option 3(C)
40	Option 1(A)	Option 5(E)
23	Option 2(B)	Option 5(E)
27	Option 2(B)	Option 1(A)
4	Option 3(C)	Option 4(D)
14	Option 3(C)	Option 1(A)
13	Option 4(D)	Option 2(B)
18	Option 4(D)	Option 3(C)
24	Option 4(D)	Option 3(C)

Conclusion

Based on the results of the analysis and discussion of the quality of multiple-choice test instruments to measure critical thinking skills worthy of being tested on students who have studied the Electrical Installation Automation material. This is based on: (1) the results of content validity found that the Aiken V value of 0.82 with the category of high validity and KR 20 shows the number 0.843 which means items are close to 1.00, then the overall item reliability is in a good range (reliable) and (2) the results of descriptive analysis assisted by test analysis program (TAP) by observing the characteristics of the items seen: the number of multiple-choice instrument items spread evenly with the difficulty level of 0.560 in the medium category, the average difference in power of 0.333 can be categorised as good and there are 39 items used indicating the trickster can function well.

References

1. Akbar, M. N., H Firman, H., & Rusyati, L., Developing Science Virtual Test to Measure Students' Critical Thinking on Living Things and Environmental Sustainability Theme, IOP Conf. Series: Journal of Physics: Conf. Series 812, (2017) doi:10.1088/1742-6596/812/1/012106
2. Arifin, Z. Evaluasi Pembelajaran. Bandung : Rosda Karya (2012)
3. Azwar , S. Reliabilitas dan Validitas. Yogyakarta: Pustaka Pelajar (2012)
4. Facione, P. A., & Facione, N. C. Critical thinking for life: Valuing, measuring, and training critical thinking in all its forms. Inquiry: Critical Thinking Across the Disciplines, 28(1), 5-25 (2013)
5. Heong, Y. M., Othman, W. D., Md Yunos, J., Kiong, T. T., Hasan, R., & Mohamad, M. M. The Level of Marzano Higher Thinking Skills Among Technical Education Students. International Journal Of Social and Humnity. 1(2): 121-125. (2011)
6. Iskandar, A. Pengembangan perangkat penilaian psikomotor di sekolah menengah kejuruan (SMK). Inspiration Jurnal Teknologi Informasi Dan Komunikasi, 3(1). (2013). Retrieved from http://jurnal.akba.ac.id/index.php/ins piration/article/view/30
7. Kothari, C.R. Reseach Methodology: Methods and Techniques. New Age International (P) Ltd, Publishers (2004)
8. Lei Mee Ing, Musah, M. B., Al-Hudawi, S. H. V., Tahir, L. M., & Kamil, N.M. Validity of Teacher-Made Assessment: A Table of Specification Approach, Asian Social Science; 11(5); 193-200 (2015) doi:10.5539/ass.v11n5p193
9. Mansyur, Rasyid, H. & Suratno. Asesmen pembelajaran disekolah. Panduan bagi guru dan calon guru. Yogyakarta: Pustaka Pelajar (2015)
10. Mohamad, M., Sulaiman, N. L., Chee Sern, L. & Salleh, K. M. Measuring the Validity and Reliability of Research Instruments, 4th World Congress on Technical and Vocational Education and Training (WoCTVET), Procedia - Social and Behavioral Sciences 204:164-171. (2015) DOI:10.1016/j.sbspro.2015.08.129
11. Moore, Betsy dan Stanley, Todd. Critical Thinking and Formative Assesments. Routledge (2013)
12. Retnawati, H. Teori respon butir dan penerapannya. Yogyakarta: Nuha Medika (2014)
13. Salvia, J., Ysseldyke, J. E., Bolt, S., Assessment in special and inclusive education. Belmont, CA USA:Wadsworth (2010)

14. Tiruneh, D. T., De Cock, M., Weldeslassie, A. G., Elen, J., & Janssen, R. Measuring critical thinking in physics: Development and validation of a critical thinking test in electricity and magnetism. International Journal of Science and Mathematics Education, 15(4), 663-682 (2017)
15. Tiruneh, D.T., De Cock, M., & Elen, J. Designing Learning Environments for Critical Thinking: Examining Effective Instructional Approaches, *Int. J. Sci. Math. Educ.*, 16(6), pp. 1065–1089 (2018)
16. Von Davier, A. A., Statistical models for test equating, scaling and linking. Princeton, New Jersey USA: Springer (2011)