Analysis of HOTS Type Multiple-choice Test Items on Learning Automotive Electrical Systems in SMK Dharma Bahari Surabaya

Soeryanto¹, I M Arsana¹, R S Hidayatullah¹, and S R Ariyanto²

¹ Department of Mechanical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, East Java, Indonesia.
² Technology and Vocational Education, Postgraduate, Universitas Negeri Surabaya, East Java, Indonesia.

Abstract. This item analysis study aims to describe the quality of HOTS type multiple-choice test items who compiled for the learning automotive electrical systems. The type of this research is document analysis. Analysis of the data used in this study is about the content validity, reliability, level of difficulty, differentiating power, and the effectiveness of detractors. The result showed that: (1) the validity of the contents of the questions are in the valid category with a percentage of 55%; (2) reliability is in the high category with a coefficient value of 0.79 or a percentage of 79%; (3) in the level of difficulty the questions are more dominant in the easy category with a percentage of 53%; (4) the differentiating power of the questions is more dominant in the good category with a percentage of 30%; and (5) most of the questions have deceptive effectiveness that works very well with a percentage of 32%.

Keyword: item analysis; content validity; reliability; the level of difficulty; differentiating power; and the level of difficulty.

1. Introduction
In Indonesia, education quality becomes very important to be improved. Education becomes an essential sector in developing nation's civilization because the ability and the character of a person can be developed by doing a series of educational activities. In this case, national education is designed to form faith, pious, noble, knowledgeable, skilled, independent, and responsible learners.

The higher the quality of education that is targeted by the government shows that the government is seriously paying attention to the education sector. However, the quality of education is influenced by many factors, one of them is the teacher’s role in the learning process. As mandated by the Government Regulation No. 19 of 2017, has been explained that the essence of professional teacher's task is not only to teach, but also educating, guiding, directing, training, assessing, and evaluating the students.

By the explanation of the government's regulation above, we can know that except teaching and educating, at the end of the course learning the teacher should evaluate the learners. It aims to find out how the development of learners during the learning process [1]. Evaluation is the process of collection, analysis, and interpretation of information about teaching and learning for making a decision [2].

To obtain the learning evaluation’s results, of course, having the instrument that can measure the accuracy in decision making is needed. Generally, the instrument that usually used is a test. In order to know the quality of a test, the analysis on each item is needed [3]. There are several ways that
can be done to analyze the items, including the analysis of the validity, reliability, level of difficulty, distinguishing features and effectiveness of detractors [4].

Based on the results of interviews conducted especially to the Light Vehicle Engineering teachers of SMK Dharma Bahari Surabaya it is known that the questions are used to do the Final Examination Final generally have never been analyzed. So that if we take a look in terms of the characteristics quality, the item’s quality is not yet known. In drawing up the question, generally, the validity of the items still tend to be neglected, for both multiple-choice questions and problem descriptions.

Referring to these problems, the authors are interested in creating 40 multiple-choice questions as the test items in Electric Light Vehicles Maintenance subject. The question’s characteristics will be analyzed both in terms of validity, reliability, level of difficulty, distinguishing features, and effectiveness of detractors (distractor). It is important to do so that the instrument used to present information is accurate and useful for teachers and students.

2. Method

2.1. Types of research
This research is the documentation analysis. This study aims to determine whether the question items of the semester final exams in learning automotive electrical systems have been qualified its validity, reliability, different power, and level of difficulty.

2.2. Subject and Object Research
This research subject is the 30 students of XI TKR 5 class of SMK Dharma Bahari Surabaya. The object of this study is the odd semester final exams in the school year 2018/2019 in learning automotive electrical systems.

2.3. Data collection technique
Data collection techniques in this study are based on a documentation technique. The documentation used is in the form of syllabus, the odd semester final exam, key answer and answer sheets throughout the test results of XI TKR 5 class student of learning automotive electrical systems in SMK Dharma Bahari Surabaya.

3. Data Analysis Technique

3.1. Validity
Validity test is used to determine the validity of items. The formula used to determine the level of validity (item) is the product moment correlation with rough numbers, as follows [5].

\[
    r_{xy} = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N(\sum X^2) - (\sum X)^2][N(\sum Y^2) - (\sum Y)^2]}}
\]  

Information:

\( r_{xy} \) = coefficient of the validity of the items
\( X \) = score item
\( Y \) = total score
\( N \) = number of respondents

3.2. Reliability
Reliability test is used to determine the extent to which assessment will produce the reliable results consistently [6], the Spearman-Brown formula is used to calculate the reliability of the formula used, as in the formula below[7].

\[
    r_{xx} = \frac{2 \times r_{xy}}{1 + r_{xy}}
\]
Information:
r_{xx} = \text{the reliability of the whole item}
\ r_{xy} = \text{correlation product moment between different parts of the odd and even}

Criteria:

- <0.20 : no reliability
- 0.21 - 0.40 : low reliability
- 0.41 - 0.70 : moderate reliability
- 0.71 - 0.90 : high reliability
- > 0.90 : very high reliability

3.3. The difficulty level

Difficulty index is used to determine the difficulty of each item. The formula used to determine the difficulty index as follows [8].

\[ P = \frac{B}{JS} \]  \hspace{1cm} (3)

Information:
P = \text{trouble Index}
B = \text{number of students who answered the question correctly}
JS = \text{the total number of student participants test}

In determining the interpretation of the difficulties level results, the criteria that can be used is as follows.
- 0.00 - 0.30 : difficult
- 0.31 - 0.70 : medium
- 0.71 - 1.00 : easy

3.4. Differentiating power

Differentiating power generally used to indicate the success or the failure of a problem in differentiating students with high abilities and students with the low ability [9]. The formula to find the differentiating power as follows [10].

\[ DP = \frac{BA - BB}{NA} \times 100\% \]  \hspace{1cm} (4)

Information:
DP: index of distinguishing certain items (one item)
BA: the number of students over the group who answered correctly
BB: the number of correct answers in the bottom group
NA: the number of students in one of the groups A or B

In determining the distinguishing interpretation of the results of calculation, the criteria that can be used is as follows.
- Negative - 9% : very bad
- 10% - 19% : bad, should be discarded
- 20% - 29% : somewhat good, would likely need to be revised
- 30% - 40% : good
- 50% and above : Excellent

3.5. Effectiveness Detractors

Effectiveness detractors are the ability of a problem in distinguishing between high-ability students and low ability students, Detractors index is calculated by the following formula.
\[ IP = \frac{P}{(N - B)/(n - 1)} \times 100\% \] (5)

Information:
IP : detractors index
P : the number of students who were choosing detractors
N : total students who take the test
B : the number of students who answered correctly
n : the number of alternative answers (option)
l : fixed number

In determining the effectiveness detractors of the calculation result interpretation, the criteria that can be used are as follows,
Very Good : 76% - 125% (approaching 100%)
Good : 51% - 75% (126% - 150%)
Less good : 26% - 50% (151% - 175%)
Bad : 0% - 25% (176% - 200%)
Very bad : more than 200%

4. Results and Discussion
4.1. Validity
The question’s validity is calculated by the formula product moment correlation. The results of calculations are consulted with \( r_{table} \) product moment at the 5% significance level. The total number of students in XI class of SMK Dharma Bahari Surabaya who take the test is 30 students. So that \( N = 30 \), which indicates 0.361 value of \( r_{table} \) product moment. The results of items validity results on the test item refer to the \( r_{xy} > 0.361 \), which means valid and \( r_{xy} < 0.361 \), which means invalid. The question’s distribution is based on the validity index as follows.

| No. | Validity Index | No. Items | Total | Percentage |
|-----|----------------|-----------|-------|------------|
| 1.  | Item Valid     | 3, 5, 6, 10, 11, 12, 15, 16, 17, 18, 19, 20, 22, 24, 26, 29, 30, 32, 34, 36, 39, 40 | 22     | 55%        |
|     | \( (r_{xy} \geq 0.361) \) |                                                     |       |            |
| 2.  | Item Invalid   | 1, 2, 4, 7, 8, 9, 13, 14, 21, 23, 25, 27, 28, 31, 33, 35, 37, 38 | 18     | 45%        |
|     | \( (r_{xy} < 0.361) \) |                                                     |       |            |
|     | Total Amount   |                                                      | 40    | 100%       |

If the validity results listed in Table 1 is presented in the form of a pie chart, it will look like in Figure 1.
Based on table 1 and figure 1 we can know that the "Valid" category have a percentage of 55%, while the category of "Invalid" has a percentage of 45% where it is demonstrated that the question of which fall into the category "Invalid" only by 45%. Thus it can be stated that 55% matter stated "Valid" of the whole learning automotive electrical systems TKR Class XI deserves to be tested.

4.2. Reliability
Reliability testing is done by grouping the total odd numbered items score as parts of the first and the even-numbered group's items score as parts of a second. After the correlation coefficient is obtained, then can coefficient reliability with Spearman-Brown formula be calculated.

Based on the reliability calculation is obtained by 0.79. These results indicate that multiple-choice questions in learning automotive electrical systems. Class XI TKR has reliability with "High" category for $0.71 < r_{xx} < 0.90$.

Based on the description above, it can be stated that multiple-choice questions in learning automotive electrical systems XI TKR class are eligible to be used as a measuring tool. When the question was tested repeatedly with the same subject, the results obtained remain the same and the high correlation. This happens because the reliability of the resulting fall into the "High" category so that the question has the consistency and reliability when used as a measuring tool.

4.3. The difficulty level
The acquisition of the measurement results, the distribution of question-based on the level of difficulty can be seen in Table 2.

| No. | The Difficulty Level | No. Items | Total | Percentage |
|-----|----------------------|-----------|-------|------------|
| 1.  | Easy (0.70 – 1)      | 2, 5, 8, 17, 23, 26, 28, 31, 34, 37, 40 | 11    | 28%        |
|     |                      | 4, 6, 7, 9, 11, 12, 14, 16, 19       |       |            |
| 2.  | Medium (0.3 – 0.69)  | 21, 22, 25, 27, 29, 30, 32, 33, 35, 36, 38, 39 | 21    | 53%        |
| 3.  | Hard (0.00 – 0.29)   | 1, 3, 10, 13, 15, 18, 20, 24          | 8     | 20%        |
|-----|----------------------|-----------|-------|------------|
|     | Total Amount         | 40        |       | 100%       |
If the result of the difficulty level listed in Table two is presented in the form of a pie chart, it will look like in Figure 2.

![Figure 2. Pie Charts Item Distribution Based on Level of difficulty](image)

Based on Table 2 and Figure 2 shows that from the total multiple-choice questions in learning automotive electrical systems XI TKR class, the question with "Medium" category has a percentage of 53% of 21 questions. Meanwhile, the issue with the "Difficult" category has a percentage of 20% of 8 questions and problems with the "Easy" category has a 27% total of 11 questions.

4.4. Differentiating power
In this matter, there are five groups that are used to determine the difference between each item. As for the distribution of matter based on different power index, as shown in Table 3.

| No. | Different Power | No. Items | Total | Percentage |
|-----|-----------------|-----------|-------|------------|
| 1.  | Very good       | 11, 16    | 2     | 5%         |
| 2.  | Good            | 3, 6, 12, 18, 22, 24, 29, 30, 32, 33, 38, 40 | 12 | 30%       |
| 3.  | Enough          | 5, 10, 15, 19, 20, 25, 31, 34, 35, 37, 39 | 11 | 27%       |
| 4.  | Bad             | 1, 4, 7, 8, 9, 13, 17, 26, 34, 35, 37, 39 | 9 | 23%       |
| 5.  | Item Not Good   | 2, 14, 21, 23, 27, 28 | 6 | 15%       |
|     | **Total Amount** | **40** | **100%** |          |

If the results of differentiating power listed in Table 3 are presented in the form of a pie chart, it will look like in Figure 3.

![Figure 3. Pie Charts Item Distribution Based on Differentiating power](image)
Based on Table 3 and Figure 3 shows that of the total multiple-choice questions in learning automotive electrical systems XI TKR class, in the "Good" category has a significant considerable contribution which has a percentage of 30%. Meanwhile, the issue with the "Enough" category has a percentage of 27%, in the "bad" category has a percentage of 23% and the question in the "Problem Not Good" category has a percentage of 15%. Then in the "Very Good" category has the lowest percentage, which is equal to 5%.

4.5. Effectiveness detractors
Detractors can be considered work as well if at least chosen by 5% of the total respondents, The calculation of the effectiveness detractors obtained by the number of choice answers are then multiplied by the total number of items. However, the key answer is not included in the calculation so that the total number become120 detractors as for the distribution of matter based on different power index in Table 4.

| No. | Detractors  | Total | Percentage |
|-----|-------------|-------|------------|
| 1.  | Very good   | 38    | 32%        |
| 2.  | good        | 35    | 29%        |
| 3.  | Enough      | 19    | 16%        |
| 4.  | bad         | 20    | 17%        |
| 5.  | Very bad    | 8     | 7%         |
|     | Total Amount| 120   | 100%       |

If the result of the power detractors listed in Table 4 is presented in the form of a pie chart, it will look like in Figure 4.

Based on Table 4 and Figure 4, we can know the percentage of effectiveness detractors. The above results indicate that from the 120 detractors of 40 items, 32% detractors included in "Very Good" categories, 29% detractors included "good" category, 17% detractors included in the "bad" category, 16% to detractors with "Not Good" category and the last was 6% for detractors in "Very bad" category.

5. Conclusions
Based on the test results and analysis contained in the discussion, it can be concluded that based on the results of the validation of the experts on the test items for maintenance of the electrical system of light vehicles get an average value of 3.37 with a very valid category. The multiple-choice questions are included in the valid category with a percentage of 55%, while the reliability is 0.79 or
79%. The items have a difficulty level with difficult criteria of 20%, moderate by 53% and easy by 27%. These results indicate that the problem with easy difficulty is more dominant by 53%. Then the items have distinguishing features with very good criteria of 5%, good by 30%, enough by 27%, bad by 23% and bad questions by 15%. These results indicate that most of the questions have a good difference with a percentage of 30%. In addition, the item has the effectiveness of deception by which the criteria are very good at 32%, good by 29%, bad by 16%, bad by 17% and very bad by 6%. These results indicate that the majority of questions have the effectiveness of fraudsters that function very well with a percentage of 32%

6. References

[1] Birgin O and Baki A 2007 The Use of Portfolio to Assess Student’s Performance J. Turkish Sci. Educ. 4 75–90
[2] Jabbarifar T 2009 The Importance of Classroom Assessment and Evaluation in Educational System International Conference of Teaching and Learning pp 1–9
[3] Dibattista D and Kurzawa L 2011 Examination of the Quality of Multiple-choice Items on Classroom Tests Can. J. Scholarsh. Teach. Learn. 2 1–23
[4] Khoshaim H B and Rashid S 2016 Assessment of the Assessment Tool: Analysis of Items in a Non-MCQ Mathematics Exam Int. J. Instr. 9 119–32
[5] Kusmaryono I, Suyitno H, Dwijanto D and Karomah N 2017 Analysis instrument test on mathematical power the material geometry of space flat side for grade 8 AIP Conf. Proc. 1867
[6] Arends R I and Kilcher A 2010 Teaching for student learning: Becoming an accomplished teacher ed L Akers, A Sharp, S Jones, C Watson and B Tamsin (New York: Taylor & Francis)
[7] Mehrens W A and Lehmann I J 1973 Measurement and evaluation in education and psychology ed J-A Weaver and H Lockwood (New York: Hold, Rinehart and Winston, Inc)
[8] Brown H D 2003 Language Assessment Principles and Classroom practices ed K Sherak, N Frantz and N McCann (San Francisco: Longman, Inc)
[9] Crocker L and Alglna J 2008 Introduction to classical and modern test theory ed M B M Staudt, M Stranz and L Stefano (New York: Holt, Rinehart and Winston, Inc)
[10] Boopathiraj C and Chellamani K 2013 Analysis of Test Items on Difficulty Level and Discrimination Index in the Test for Research in Education Int. J. Soc. Sci. Interdiscip. Res. 2 189–93