

Item Analysis of Multiple Choice: Tool for Assessing Knowledge of Potato and Pasta Dishes in Vocational Schools

Niki Fadilla^{1,*} Emi Susilowati² Sri Handajani² Ekohariadi³
Savira U. Hasanah⁴ Muhammad Y. Pratama⁵

¹ Technology and Vocational, Postgraduate, Universitas Negeri Surabaya, Surabaya, Indonesia

² Department of Home Economic, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia

³ Department of Electrical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia

⁴ Department of Linguistics, Faculty of Humanities, Universitas Airlangga, Surabaya, Indonesia

⁵ Department of Mechanical Engineering, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia

*Corresponding author. Email: nikifadilla@unesa.ac.id

ABSTRACT

Multiple-choice test is one of the most widely and popularly used tools to evaluate students' competence and knowledge. This type of test mainly used at various levels considering its efficiency, reliability, and the easiness in the standardization process when compared to other types of tests. But unfortunately, until now there are still many teachers who do not care about the importance of doing item analysis. The main objective of this paper is to produce a multiple-choice test on the competence of potato and pasta dishes with high quality so that it can be maximized to assess the knowledge of vocational school students. This research is a descriptive quantitative study, with 34 students as the data samples. The data of this research are the students' responses to the final exam which consists of 50 multiple-choice items with five choices for each item. The findings of this study include as many as 40 questions that fall into the proper category. The difficulty index also shows that sequentially 21 questions include in easy category, 27 questions are in the medium category, while 2 questions are in the difficult category. Meanwhile, the items also have a good discrimination index were considered as the bad and very bad categories only 10 students are selected. Furthermore, from 150 distractors for distractor efficiency, there are at least 29 answer choices in the bad category and 10 questions in the very bad category.

Keywords: *Item analysis, Multiple-choice, Vocational schools, Validity index, Reliability index.*

1. INTRODUCTION

Multiple-choice test is widely used for evaluation at all education levels from basic education, secondary education, to higher education [1]. This test is basically designed with the aim of measuring the extent to which students are able to increase their knowledge after going through a series of learning processes [2]. However, if the test is prepared carefully, it can be used to assess other abilities such as critical thinking or problem-solving [3]. Composing good questions (valid and reliable) is certainly not an easy task. The quality review of the test must be carried out carefully and thoroughly on each item. A multiple-choice test is a set of questions that are used to test a certain knowledge or competence. Each item consists a set of answer choices that must be

completed [4]. Generally, the answer choices are approximately 3 to 5 answers. One of several choices is the correct answer or often called the answer key, while the wrong answer is used as a distractor[5]. In multiple-choice test, students who can choose the correct answer get a score (1), while if they are wrong they get (0) [6]. Thus, students are expected to be able to be careful in determining the answers during the test.

In general, multiple-choice tests are widely used at various levels because of efficiency, reliability, and the easiness of standardization when compared to other types of tests. In addition, this type of test also can improve Higher Order Thinking Skills (HOTS). If it is arranged well, the choice questions not only assess the knowledge memory (C1), but also understanding (C2), application (C3), analysis (C4), synthesis (C5), and evaluation (C6)

according to the framework of Bloom's Taxonomy and the Pyramid Miller [7]. Building good multiple-choice questions is a challenge and of course time-consuming, but when compared to other test methods, multiple-choice questions is more objective and have minimal errors [8]. The analysis process for each item must be carried out starting from the validity index, reliability, difficulty index, discrimination index, to distractor efficiency tests.

Basically, the implementation of item analysis before the test is very important. The goal is that there is no doubt about the ability of the test in measuring students' competence or knowledge [9]. In addition, the results of item analysis can be used as a feedback regarding whether the questions that have been prepared are suitable or need to be improved [10]. But unfortunately, until now there are still many teachers who are not aware of the importance of doing item analysis [11]. Especially in Indonesia, until now most teachers are still compiling questions only referring to the basic competencies that have been carried out by the government without paying attention to whether the questions are worth testing on students or not. The problem is that it only covers the realm of knowledge (C1) and understanding (C2). Whereas in the 21st-century learning era, students should be directed to improve their abilities at the level of application (C3), analysis (C4), synthesis (C5), and evaluation (C6).

Furthermore, Trilling & Fadel emphasized that students are able to improve 4C skills which include: (1) critical thinking and problem-solving; (2) creativity; (3) communication skills; and (4) ability to work collaboratively [12]. Therefore, the main objective of this paper is to produce a multiple-choice test on the competence of potato and pasta dishes with high quality so that it can be maximized to assess the vocational school students' knowledge. The quality of this test is measured based on the results of the validity index, reliability, difficulty index, discrimination index, and distractor efficiency test results. With this method, the teacher can detect the items that are less feasible so that it can be decided whether the item should be revised or not.

2. METHODS

This research is quantitative descriptive research conducted at YPM 2 Taman Sidoarjo. The population in this study were students of class XI catering, while the research sample was selected using the purposive sampling technique. The results are converging on the decision that the sample of this study was 34 students of class XI Catering II. The data of this research are students' responses to the final exam which consists of 50 multiple choice questions that have five choices for each item. The data were analyzed quantitatively using the help of Microsoft Excel 2016.

2.1. Validity Index

The validity test was conducted to analyze how many items were in the valid and invalid categories. It is calculated using the product-moment correlation formula with rough numbers as follows [13].

$$r_{xy} = \frac{N(\sum XY) - (\sum X)(\sum Y)}{\sqrt{[N(\sum X^2) - (\sum X)^2][N(\sum Y^2) - (\sum Y)^2]}} \quad (1)$$

- r_{xy} : Coefficient of the validity of the items
- Y : Total score
- X : Score item
- N : Number of respondents

2.2. Reliability

A reliability test is used to measure the items' consistency [14]. Reliability relates to the stability or consistency of scores over time or across raters. Considering that reliability is related to scores, not to the people [15]. In this study, the reliability test was carried out through the Cronbach's Alpha technique assisted by SPSS 16 software.

2.3. Difficulty index

The difficulty index is the percentage of each item that falls into the easy, medium, and difficult categories. It is ranged between 0% and 100% [16]. It is calculated using a formula as follows:

$$\text{Difficulty Index} = \frac{\text{Students with correct answers}}{\text{Number of students taking the test}} \quad (2)$$

2.4. Discrimination Index

Discrimination Index is the ability of an item to discriminate between students who are smart or highly capable and students who are less intelligent or have low abilities [16]. It is calculated using a formula as follows:

$$\text{Discrimination Index} = \frac{CU - CL}{N} \times 100\% \quad (3)$$

- CU : The number of correct answers in the upper group
- CL : The number of correct answers in the lower group
- N : Number of students taking the test

2.5. Distractor efficiency

Distractor efficiency (DE) is determined based on the number of Non-functional distractors (NFD) selected in the item. NFD is a distractor selected at least less than 5% of students in an item [7]. DE generally ranges from 0 to 100%. It is calculated using a formula follows:

$$DE = \frac{NFD}{(NST - SWA)/(NAO - 1)} \times 100\% \quad (3)$$

- DE : Distractor efficiency

- SWA : Students with correct answers
- NFD : Non-functional distractors
- NAO : Number of answer options
- NST : Number of students taking the test

3. RESULTS AND DISCUSSION

3.1 Validity Index

Table 1. Validity index test results

Item Statistic	Interpretation	Total item (%)
$Y_{pbi} \geq 0.339$	Valid	40 (80%)
$Y_{pbi} \leq 0.339$	Not valid	10 (20%)

Item validity is an item that is considered as valid if it has great support for the total score. The score on the item causes the total score to be high or low [17]. The items were tested with a significance level of 5% with a total (N) of 34 students, so the r_{table} product-moment value was 0.339. The item is d valid if $Y_{pbi} > 0.339$ and the item is invalid if $Y_{pbi} < 0.339$. From the validity test result as shown in Table 1 and Figure 1a, it is known that out of a total of 50 multiple choice questions there are 40 questions that fall into the "Valid" category with a percentage of 80%, while there are 10 questions that fall into the "Invalid" category with a percentage of 20%. Thus, it is stated that there are 40 questions suitable for use in the competence of potato and pasta dishes.

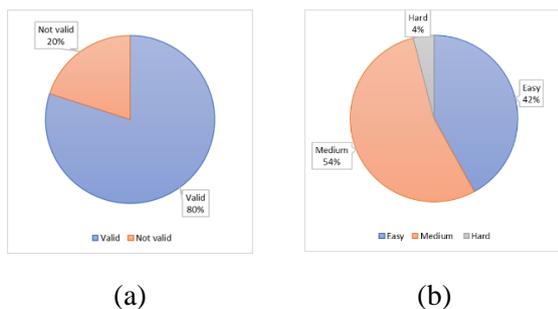


Figure 1 Pie chart (a) validity index and (b) difficulty index.

3.2 Reliability

Testing the reliability of the test was carried out using the Cronbach's Alpha technique. Reliable is related to constancy. No matter how many items are tested, they have almost the same value [18]. The reliability test results can be seen in Table 2.

Table 2. Item reliability test results

Reliability Statistics	
Cronbach's Alpha	N of items
0,833	50

Based on the table above, the results obtained are 0.833. Thus, the items can be categorized as "special" which means that the multiple-choice test of competence in potato and pasta dishes is feasible and reliable.

3.3 Difficulty index

Table 3. Difficulty index test results

Item Statistic	Interpretation	Total item (%)
0.70 – 1	Easy	21 (42%)
0.3 – 0.69	Medium	27 (54%)
0.00 – 0.29	Hard	2 (4%)

The difficulty level item is the proportion between the number of students who answered the item correctly and the total of students. This means that the more students who answer the item correctly, the greater the index of difficulty level, which means the easier the item is [19]. The level of difficulty is divided into 3 groups - easy, medium, and difficult. The test item results in difficulty level can be seen in Table 3 and Figure 1b, wherefrom the table, it is known that from a total of 50 multiple-choice items, there are 21 questions that fall into the "Easy" category with a percentage of 42%, while there are 27 questions that fall into the "Medium" category. ” with a percentage of 54% and there are 2 questions that fall into the “Difficult” category with a percentage of 4%.

3.4 Discrimination Index

Table 4. Discrimination index test results

Item Statistic	Interpretation	Total item (%)
$0,700 \leq DI \leq 1,000$	Very good	5 (10%)
$0,400 \leq DI < 0,700$	Good	20 (40%)
$0,200 \leq DI < 0,400$	Pretty good	15 (30%)
$0,00 \leq DI < 0,200$	Bad	2 (4%)
$DI < 0$	Very bad	8 (16%)
$0,700 \leq DI \leq 1,000$	Very good	5 (10%)

Item discrimination is the ability of an item to distinguish groups in the aspects measured according to the differences in the group [20]. In this study, the discriminatory test of the items was divided into 5 categories, (1) very good; (2) good; (3) sufficient; (4)

bad; and (5) very bad. The discrimination index analysis was carried out using Microsoft Excel 2016 with the results as shown in Table 4 and Figure 2a. Based on the discrimination index data in the table, it is known that out of a total of 50 multiple choice questions, there are 5 questions that fall into the "Very good" category with a percentage of 10%. , 20 questions fall into the "Good" category with a percentage of 40%, 15 questions fall into the "Enough" category with a percentage of 15%, 2 questions fall into the "Bad" category with a percentage of 2% and 8 questions fall into the "Very good" category. bad" with a percentage of 16%.

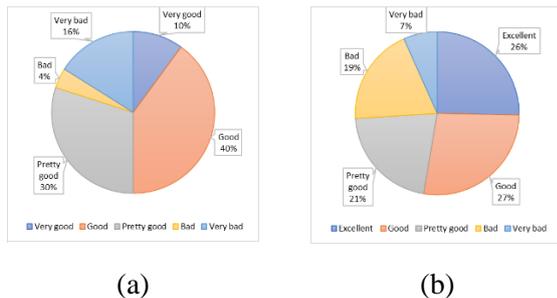


Figure 2 Pie chart (a) discrimination index and (b) distractor efficiency.

3.5 Distractor efficiency

Table 5. Distractor index test results

Item Statistic	Interpretation	Total item (%)
50% and above	Excellent	38 (25%)
30% - 40%	Good	41 (27%)
20% - 29%	Pretty good	32 (21%)
10% - 19%	Bad	29 (19%)
Negative - 9%	Very bad	10 (7%)
50% and above	Excellent	38 (25%)

From the answer questions' pattern, it can be determined whether the distractor functions work well or not. Distractors who are not chosen at all by the respondents can be considered that the distractors are bad, too glaringly misleading. On the other hand, a distractor is considered to work well if it has a great appeal to test takers who do not understand the concept or master the material [21]. The results of the distractor efficiency analysis in this study (Table 5 and Figure 2b) showed that from a total of 150 distractors there were 38 questions in the "Very Good" category with a percentage of 25% distractors, 41 questions in the "Good" category with a percentage of 27% distractors, 32 questions in the distractor efficiency category. "Not Good" with a percentage of 21% distractors, 29 questions fall into the "Bad" category with a percentage of 19% distractors and 10 questions fall into the "Very Bad" category with a percentage of 7% distractors.

4. CONCLUSION

In this paper, multiple-choice test analysis was conducted to measure the extent to which the items were appropriate to be used to measure students' knowledge of potato and pasta dish competence. Based on the validity of the items, as many as 40 questions fall into the proper category. The difficulty index also shows that sequentially 21 questions are in the easy category, 27 questions are in the medium category, while 2 questions are in the difficult category. Meanwhile, the items also have a good discrimination index where for the bad and very bad categories only 10 students are selected. Furthermore, from a total of 150 distractors for distractor efficiency, there are at least 29 answer choices in the bad category and 10 questions in the very bad category.

AUTHORS CONTRIBUTION

All authors conceived and designed this study. All authors contributed to the process of revising the manuscript, and at the end all authors have approved the final version of this manuscript.

REFERENCES

- [1] G.T.L. Brown, H.H.A. Abdulnabi, Evaluating the Quality of Higher Education Instructor-Constructed Multiple-Choice Tests: Impact on Student Grades Front. Educ. 2, 2(24) (2017) 1-12 DOI: <https://doi.org/10.3389/FEDUC.2017.00024>
- [2] M. Amo-Salas, M.d.M. Arroyo-Jimenez, D. Bustos-Escribano, E. Fairén-Jiménez, J. López-Fidalgo, New Indices for Refining Multiple-choice Questions J. Probab. Stat. 2014(2014) 1–8 DOI: <https://doi.org/10.1155/2014/240263>
- [3] S.R. Ariyanto, I.W.P. Lestari, S.U. Hasanah, L. Rahmah, D.V. Purwanto, Problem Based Learning dan Argumentation Sebagai Solusi dalam Meningkatkan Kemampuan Berpikir Kritis Siswa SMK J. Kependidikan J. Has. Penelit. dan Kaji. Kepustakaan di Bid. Pendidikan, Pengajaran dan Pembelajaran 6(2) (2020) 197 DOI: <https://doi.org/10.33394/jk.v6i2.2522>
- [4] C.E. Brasil, B.A. Couch, Multiple-true-false questions reveal more thoroughly the complexity of student thinking than multiple-choice questions: a Bayesian item response model comparison Int. J. STEM Educ. 6(16) (2019) 1-17 DOI: <https://doi.org/10.1186/s40594-019-0169-0>
- [5] J. Shin, Q. Guo, M.J. Gierl, Multiple-Choice Item Distractor Development Using Topic Modeling Approaches Front. Psychol. 10(825) (2019) 1-14 DOI: <https://doi.org/10.3389/fpsyg.2019.00825>

- [6] J. Vanderroost, R. Janssen, J. Eggermont, R. Callens, T. De Laet, Elimination testing with adapted scoring reduces guessing and anxiety in multiple-choice assessments, but does not increase grade average in comparison with negative marking ed M B Steinborn PLoS One, 13(10) (2018) e0203931 DOI: 10.1371/journal.pone.0203931
- [7] S. Harti, A.K. Mahapatra, S.K. Gupta, T. Nesari, All India AYUSH post graduate entrance exam 2019 – AYURVEDA MCQ item analysis J. Ayurveda Integr. Med. 12(2) (2021) 356-358 DOI: 10.1016/j.jaim.2021.01.013
- [8] Kumar D, Jaipurkar R, Shekhar A, Sikri G and Srinivas V 2021 Item analysis of multiple-choice questions: A quality assurance test for an assessment tool Med. J. Armed Forces India 77 S85–9
- [9] D. Kumar, R. Jaipurkar, A. Shekhar, G Sikri, V. Srinivas, Item analysis of multiple choice questions: A quality assurance test for an assessment tool, Med. J. 77(Suppl 1) (2018) S85-S89 DOI: 10.1016/j.mjafi.2020.11.007
- [10] Y.R. Karkal, G.S. Kundapur, Item analysis of multiple-choice questions of undergraduate pharmacology examinations in an International Medical School in India J. Dr. NTR Univ. Heal. Sci. 5(3) (2016) 183-186 URL: <https://www.jdrntruhs.org/article.asp?issn=2277-8632;year=2016;volume=5;issue=3;spage=183;epage=186;aualst=Karkal;type=0>
- [11] J.A. Odukoya, O. Adekeye, A.O. Igbinoaba, A. Afolabi, Item analysis of university-wide multiple-choice objective examinations: the experience of a Nigerian private university Qual. Quant. 52(3) (2018) 983–97 DOI: 10.1007/s11135-017-0499-2
- [12] B, Trilling, C. Fadel, 21st century skills: learning for life in our times, San Francisco: Jossey-Bass, 2009.
- [13] I. Kusmaryono, H. Suyitno, Dwijanto, N. Karomah, Analysis instrument test on mathematical power the material geometry of space flat side for grade 8 AIP Conf. Proc. 1867 (2017) 020013 DOI: <https://doi.org/10.1063/1.4994416>
- [14] R.I. Arends, A. Kilcher, Teaching for student learning: Becoming an accomplished teacher ed L Akers, A Sharp, S Jones, C Watson and B Tamsin, New York: Taylor & Francis, 2010.
- [15] O. Bolarinwa, Principles and methods of validity and reliability testing of questionnaires used in social and health science researches Niger. Postgrad. Med. J. 22(4) (2015) 195-201 DOI: 10.4103/1117-1936.173959
- [16] M. Kaur, S. Singla, R. Mahajan, Item analysis of in use multiple-choice questions in pharmacology Int. J. Appl. Basic Med. Res. 6(3) (2016) 170-3 DOI: 10.4103/2229-516X.186965.
- [17] W. Warju, S.R. Ariyanto, S. Soeryanto, R.A. Trisna, Analisis Kualitas Butir Soal Tipe Hots pada Kompetensi Sistem Rem di Sekolah Menengah Kejuruan J. Pendidik. Teknol. dan Kejur. 17(1) (2020) 95-104 DOI: <http://dx.doi.org/10.23887/jptk-undiksha.v17i1.22914>
- [18] Soeryanto, I.M. Arsana, R.S. Hidayatullah, S.R. Ariyanto, Analysis of HOTS Type Multiple-choice Test Items on Learning Automotive Electrical Systems in SMK Dharma Bahari Surabaya J. Phys. Conf. Ser. 1569 (2019) 032046 DOI: 10.1088/1742-6596/1569/3/032046
- [19] S. Azwar, Tes Prestasi: Fungsi dan Pengembangan Pengukuran Prestasi Belajar ed Pustaka Pelajar, Yogyakarta: Pustaka Pelajar Offset, 2015.
- [20] K. Quaigrain, A.K. Arhin, Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation ed S King Fai Hui Cogent Educ. 4(1) (2017) 1–11 DOI: 10.1080/2331186X.2017.1301013
- [21] M. Tweed, Adding to the debate on the numbers of options for MCQs: the case for not being limited to MCQs with three, four or five options BMC Med. Educ. 19(354) (2019) 1-4 DOI: 10.1186/s12909-019-1801-x