Analysis of Test Items in Motorcycle Engine Maintenance for Vocational High School Students

Soeryanto Soeryanto *, Rachmad Syarifudin Hidayatullah, Dany Iman Santoso, Muhamad Febriansyah

Department of Mechanical Engineering Education, Faculty of Engineering, Universitas Negeri Surabaya, Surabaya, Indonesia
* *Corresponding author. Email: soeryanto@unesa.ac.id*

ABSTRACT
This study aims to determine the quality of motorcycle engine maintenance items for Vocational High Students (VHS) students. The item analysis is seen from various aspects including the level of validity, reliability, level of difficulty, discriminating power and the effectiveness of using distractors. This research is descriptive in nature using quantitative methods because all data or information obtained is presented in numerical form and analyzed statistically using the Item and Test Analysis (ITEMAN) program. The subjects of this study were students majoring in TBSM SMK. Data collection techniques were carried out in order to obtain question data, answer keys, and student test results. The results of the study explained that: (1) Based on the validity of the items on the verification results of the three validators (language, study and social validators), it is known that no revisions have been made to the questions. (2) Based on the reliability of the questions, an Alpha of 0.914 is included in the very high category, namely $0.80 < r \leq 1.00$. The standard measurement error is 2,863. (3) In terms of difficulty, 3 questions are classified as easy questions (6.38%), 37 items are classified as moderate questions (78.72%), and 7 items are classified as difficult questions (14.89%). (4) In terms of discrimination, there are 10 items (21.27%) with bad discrimination, 11 items (23.40%) are unsatisfactory, 1 item is satisfactory (2.12%), and 25 items are satisfactory (53.19%). (5) Judging from the effect of the use of distractors, if the value of distractors A, B, D, and E is less than 5%, it means that the distractor is not good enough (rejected) and must be replaced. In addition, out of 47 class XI student test results, there were 12 questions that needed to be revised, while 35 questions were maintained.

Keywords: *Item analysis, Validity, Reliability, Vocational students.*

1. INTRODUCTION

Education is a very important factor in fostering and developing the potential of each individual [1][2]. As stated in Article 1 of the Law of the Republic of Indonesia Number 20 of 2003 concerning the National Education System, education is a conscious and planned effort to create a learning atmosphere and learning process that enables students to actively develop their potential and acquire a religious spirit. Strength, self-control, character, wisdom, noble character and skills needed by oneself, society, nation and state. In the world of education, learning is divided into three phases, namely planning, implementation and evaluation. The assessment stage measures how well the learning objectives are achieved [3]. According to [4], assessment is an assessment activity that aims to measure the success or failure of the learning process.

Measurement involves comparing an observation with a certain standard. The measurement results are then interpreted and calculated in the assessment [5]. The assessment process involves making decisions to set goals, learning outcomes, feedback from students [6]. Meanwhile, the assessment aims to determine the level of proficiency and understanding of students in mastering the material being taught [7]. Educational evaluation is a process that cannot be separated from learning activities

A. Kusumastuti et al. (eds.), *5th Vocational Education International Conference (VEIC 2023)*,
Advances in Social Science, Education and Humanities Research 813,
https://doi.org/10.2991/978-2-38476-198-2_116

[8], because learning activities must be followed by assessment activities [9]. Evaluation aims to identify the efforts that have been made in the learning process that are carried out well or not. Evaluation is carried out to determine the success or failure of education in achieving its goals [10]. In order to achieve learning objectives, evaluation must be carried out systematically and continuously to find a good learning process [11][12]. Learning requires evaluation to find out how effective the learning activities have been implemented on students [13].

The test is an assessment technique where students are asked to complete various tasks that must be done to produce values about student behavior [14]. At the Vocational High School level, learning evaluation is carried out per semester and is held twice, namely the Semester Final Assessment and Semester Final Assessment. In line with learning in the 21st century which includes critical thinking, creative and innovative thinking, as well as communication skills and collaboration skills [15]. Critical thinking is the ability used to prioritize thinking, process and manage information so that its validity can be accounted for [16]. Thus, teachers must develop student skills that are relevant to the 21st century, one of which is critical thinking skills [17][18]. So, in its application using the Problem Based Learning (PBL) learning method and the questions used in the exam must contain HOTS (High Order Thinking Skill) questions. From the Introduction to the Schooling Environment (ISE) activity, the test items were not tested for the quality of the questions in writing, resulting in a pseudo-assessment that resulted in not measuring actual learning abilities.

Item analysis can be used as diagnostic information to determine whether students understand what they have learned and to improve the quality of the questions by correcting or eliminating invalid questions [19][20]. This study aims to determine the quality of the mid-semester assessment items which analyze the difficulty index, discriminating power index, validity and reliability of the items tested which are useful for improving the management of the implementation of learning evaluation and increasing information about students' abilities and the quality of the questions given with rasch model [21][22]. The Rasch model has similarities with the 1PL model, namely measurements that both emphasize the level of difficulty [23].

2. METHOD

This research is included in expo facto research, adopts quantitative descriptive method, and uses ITEMAN with the aim of analyzing the level of validity, reliability, difficulty level, discriminating power and deceptive validity by using the items measured. This research is descriptive quantitative in nature, because all data or information obtained is presented in the form of numbers and analyzed statistically using the Item and Test Analysis (ITEMAN) program. The subjects in this study were Motorcycle Engineering and Business (MEB) students, totaling 35 students. Data collection was obtained from questions, answer keys, and student scores. Data collection tools were given to students in the form of questions through answer sheets. There are a total of 35 questions in the form of multiple-choice questions. From the results of student responses, if you answer the question correctly, you will get a value of 1. If you answer the question incorrectly or don't even answer it, you will get a value of 0.

3. RESULTS AND DISCUSSION

Table 1. Empirical instrument analysis.

Spesification	Value	Spesification	Value
Number of examinees	35	Total Items	35
Scored Items	35	Pretest Items	0
Multiple Choice Items	35	Polytomous	0
Number of Domains	1	External Scores	No
Minimum P	0.00	Maximum P	1.00
Maximum Item Mean	0.00	Maximum Item Mean	15.00
Maximum Item Correlation	0.00	Maximum Item Correlation	1.00
ITEMAN 3.0 Header	No	Exclude Omits from Option Statistics	No
Number of ID Colums	5	ID Begins in Column	1
Responses Begin in Column	6	Omit Character	0
Not Admin Character	N	Produce Quantile Tables	Yes
Correct for Spuriousness	Yes	Produce Quantile Prlots	Yes
Save Data Matrix	No	Include Omit Codes in Matrix	N/A
Include Not Admin Codes in Matrix	N/A	Include Scaled Scores for	Total Score
Scaling Function	Standarized	Scaled Score New Mean	0.000
Scaled Score New SD	1.000	Dichotomous Classification	No
Classify Based On	N/A	Cutpoint	N/A
Low Group Label	Low	High Group Label	Hight
Data is Delimited by	N/A	Test for DIF	No
Group Status is In Column	N/A	Ability Levels for DIF	N/A
Group 1 Code	N/A	Group 2 Code	N/A

Group 1 Label	N/A	Group 2 Label	N/A

Table 1 shows that the number of test takers was 35 people, the test score was 35, the total number of 35 questions, the types of questions were 35 multiple choice questions, the number of domains (types of questions) was 1, the number of letters for identity was 5, namely ID001, while the column letters for answers starting from the 6th column, namely ID001A.

3.1. Coefficient of Meaning Reliability

Table 2. Empirical instrument analysis.

Criteria r	Information
$r \leq 0.20$	Very Low
$0.20 < r \leq 0.40$	Low
$0.40 < r \leq 0.60$	Currently
$0.60 < r \leq 0.80$	High
$0.80 < r \leq 1.00$	Very High

Reliability is a coefficient that indicates how reliable a device/measuring device is [24][25]. This means that the results are relatively stable or consistent when the instrument is used repeatedly to measure the same thing. Empirically, high or low reliability is indicated by a numerical value called the reliability coefficient [26]. The confidence factor magnitude ranges from 0 to 1. The higher the reliability number, the more consistent the measurement results. However, empirically, a reliability factor of 1 is rarely reached.

3.2. Coefficient of Meaning Reliability

Table 3. Difficulty level criteria.

Criteria P	Information
$P > 70$	Easy
$0.30 \leq P \leq 0.70$	Currently
$P < 0.30$	Hard

The level of difficulty of a question is the proposition or percentage of subjects who answer certain test items correctly. While the number that indicates the difficulty or not of the items in the test is called the index (denoted by p). and hard. The difficulty level of the item is the proportion between the number of test takers who answered the item correctly to the number of test takers [27]. This means that the more test takers who answer the item correctly, the greater the index of difficulty level, which means the easier the item is. On the other hand, the fewer test takers who answered the items correctly, the more difficult the questions were. Meanwhile, according to Hopkins in the book of Classroom Measurement and Evaluation, the level of difficulty of the items is measured by the percentage of students who answered the questions correctly. If the questions are easy, the difficulty index is higher. Questions with a p value close to 0 are very difficult questions, while questions with a p value close to 1 are very easy questions. The index of very good difficulty level is 0.3 to 0.7 [28].

3.3. Coefficient of Meaning Reliability

Table 4. Different power criteria.

Criteria	Information
0.40 – 1.00	Very Satisfactory
0.30 – 0.39	Satisfactory
0.20 – 0.29	Not Satisfying
Negatif – 0.19	Bad

Analyzing different power means studying test questions in terms of the ability of the test to distinguish students who fall into the weak/low category and the strong/high achievement category [29][30]. The discussion of the differentiating power of the items in Anates can be seen in the table of differentiating power in the percent DP column. Items with a differentiability index $\geq$ 0.30 were declared good and items with a differentiability index < 0.30 were declared not good. The different power of item items has benefits, namely to improve the quality of each item of empirical data and to find out how far each item can distinguish students' abilities, namely students who have understood or have not understood the material taught by educators.

Table 5. Statistical summary analysis.

Score	Items	Mean	SD	Min Score	Max Score	Mean P	Mean Rpbis
Score Items	35	18.257	6.532	5	30	0.522	0.387
Scaled Total	35	0.000	1.000	-2.030	1.798	-	-

Table 6. Reliability.

Score	Alpha	SEM	Split-Half (Random)	Split-Half (First-Last)	Split-Half (Odd-Even)	S-B (Random)	S-B (First-Last)	S-B (Odd-Even)
Score Items	0.884	2.228	0.800	0.756	0.755	0.889	0.861	0.861

It is explained in table 5 that there are 35 items analyzed, average 18,257, standard 6,532, minimum score 5, and maximum score 30. Based on Table 3, the level of difficulty (average P) is 0.522, and P≤0.70 indicates that the questions analyzed included medium difficulty. The Differential Power (Rpbis) in the table above is obtained from the Mean Rpbis 0.387, according to table 4 the value of Rpbis 0.30 – 0.39 means that the item has a satisfactory differential power.

The reliability value (Alpha) shown in the table above found an Alpha of 0.884 based on the Reliability Coefficient table. The value of the item being analyzed has a reliability of 0.80 <r ≤ 1.00 Very high with a standard error of measurement of 2.228.

Tabel 7. Item reliability.

No Item	P	Info	Rbpis	Info	Alpha	Info	Df>5%	Info
Item 01	0.714	easy	0.044	Very Satisfactory	0.888	Very High	does not work	Defended
Item 02	0.886	easy	0.456	Very Satisfactory	0.88	Very High	does not work	Defended
Item 03	0.714	easy	0.024	Very Satisfactory	0.888	Very High	function	Defended
Item 04	0.714	currently	0.609	Very Satisfactory	0.876	Very High	does not work	Defended
Item 05	0.057	hard	0.346	Very Satisfactory	0.882	Very High	function	Defended
Item 06	0.2	hard	0.519	Very Satisfactory	0.878	Very High	function	Defended
Item 07	0.429	currently	-0.184	Bad	0.893	Very High	function	Defended
Item 08	0.914	easy	0.081	Bad	0.885	Very High	does not work	Revisied
Item 09	0.886	easy	0.368	Satisfying	0.881	Very High	does not work	Defended
Item 10	0.029	hard	0.184	Bad	0.884	Very High	does not work	Revisied
Item 11	0.257	hard	0.469	Very Satisfactory	0.879	Very High	function	Defended
Item 12	0.457	currently	0.63	Very Satisfactory	0.875	Very High	function	Defended
Item 13	0.914	easy	0.212	Not Satisfying	0.883	Very High	does not work	Defended
Item 14	0.429	currently	0.719	Very Satisfactory	0.873	Very High	does not work	Defended
Item 15	0.457	currently	0.802	Very Satisfactory	0.871	Very High	does not work	Defended
Item 16	0.514	currently	0.702	Very Satisfactory	0.874	Very High	does not work	Defended
Item 17	0.886	easy	0.309	Satisfying	0.882	Very High	does not work	Defended
Item 18	0.686	currently	-0.33	Bad	0.895	Very High	function	Defended
Item 19	0.514	currently	0.563	Very Satisfactory	0.877	Very High	does not work	Defended
Item 20	0.429	currently	0.678	Very Satisfactory	0.874	Very High	function	Defended
Item 21	0.429	currently	0.801	Very Satisfactory	0.871	Very High	function	Defended
Item 22	0.4	currently	0.578	Very Satisfactory	0.877	Very High	function	Defended
Item 23	0.429	currently	0.78	Very Satisfactory	0.872	Very High	function	Defended
Item 24	0.886	easy	0.501	Very Satisfactory	0.879	Very High	function	Defended
Item 25	0.829	easy	0.115	Bad	0.885	Very High	function	Defended
Item 26	0.514	currently	0.763	Very Satisfactory	0.872	Very High	does not work	Defended
Item 27	0.886	easy	0.251	Not Satisfying	0.883	Very High	does not work	Defended
Item 28	0.114	hard	0.367	Satisfying	0.881	Very High	function	Defended
Item 29	0.743	easy	0.587	Very Satisfactory	0.877	Very High	does not work	Defended
Item 30	0.114	hard	0.308	Satisfying	0.882	Very High	function	Defended
Item 31	0.143	hard	0.387	Satisfying	0.881	Very High	does not work	Defended
Item 32	0.343	currently	0.292	Not Satisfying	0.883	Very High	does not work	Defended
Item 33	0.114	hard	0.006	Bad	0.886	Very High	function	Defended
Item 34	0.457	currently	0.72	Very Satisfactory	0.873	Very High	function	Defended
Item 35	0.771	easy	-0.097	Bad	0.89	Very High	does not work	Revisied

Table 7 shows that the questions with easy difficulty level are 12 item items (34.28%), in the medium category are 15 item items (42.85%) with the difficult category being 8 item items (22.85%), if we compare it with the standard distribution of questions where 30 % -40% easy, 60% -80% moderate, 30% -40% difficult, then in the category of easy questions it does not meet the standard because the easy questions are only 6.8%, for the category of medium questions when compared to the standard where the ratio is 60% - 80%, then the questions in the moderate category are in accordance with the standard, while for the questions in the difficult category, when compared with the standard, namely 30% -40%, the questions in the difficult category are not in accordance with the standard. for rpbis different power, questions with poor different power were 7 items (20%), 3 items (8.57%) unsatisfactory, 8 items (25.85%) satisfying, while for the very satisfying category 19 items (54.28%)

). For Alpha (reliability) 35 items the very high category is at 0.80 <r ≤ 1.00, the distractor factor indicates a malfunction of the distractor does not work on 18 items, namely on items 1, 02, 13, 27, 32, 08, 35, 10, 09, 17, 31, 29, 04, 14, 15, 16, 19, 26. A value of <5% means that the distractor is not good (rejected) and must be replaced, other than this from the results of testing questions on students of class XI out of 35 there are 2 questions that need to be revised while 33 questions are maintained

4. CONCLUSION

Based on the results of the analysis of the validity, reliability, discriminating power, level of difficulty, and detractor validity items, it can be concluded that the motorcycle engine maintenance instrumentation questions are very good quality questions, indicating that out of 35 items only 2 questions need to be revised.

SUGGESTION

The maker of instrument questions for the subject of motorcycle engine maintenance further increases the ability and understanding of the questions. This is because the number of very good, good, and moderate quality questions is less than the poor and very bad quality questions. Conversely, questions that are of high quality better describe the status of students' abilities than questions that are not of low quality.

REFERENCES

[1] S. Sedlacek, The role of universities in fostering sustainable development at the regional level, J. Clean. Prod., 48, 2013, pp. 74–84. DOI: 10.1016/j.jclepro.2013.01.029.

[2] A. Ferrari, R. Cachia, and Y. Punie, Innovation and Creativity in Education and Training in the EU Member States: Fostering Creative Learning and Supporting Innovative Teaching, Seville: Luxembourg, 2009.

[3] M. Erfan, M. A. Maulyda, V. R. Hidayati, F. P. Astria, and T. Ratu, Analisis kualitas soal kemampuan membedakan rangkaian seri dan paralel melalui teori tes klasik dan model rasch, Indones. J. Educ. Res. Rev., 3(1), 2020, p. 11. DOI: 10.23887/ijerr.v3i1.24080.

[4] Arikunto, Prosedur Penelitian : Suatu Pendekatan Praktik, Rineka CIpta, 2011.

[5] N. F. Zainal, Pengukuran, Assessment dan Evaluasi dalam Pembelajaran Matematika, Laplace J. Pendidik. Mat., 3(1), 2020, pp. 8–26. DOI: 10.31537/laplace.v3i1.310.

[6] M. S. Ibarra-Sáiz, G. Rodríguez-Gómez, and D. Boud, The quality of assessment tasks as a determinant of learning, Assess. Eval. High. Educ., 46(6), 2021, pp. 943–955. DOI: 10.1080/02602938.2020.1828268.

[7] M. F. Kalahatu, Persepsi peserta pelatihan dasar terhadap penggunaan quizizz sebagai metode evaluasi pembelajaran, Akademika, 10(1), 2021, pp. 163–178. DOI: 10.34005/akademika.v10i01.1228.

[8] T. Li, How formative are assessments for learning activities towards summative assessment?, Int. J. Teach. Educ., 9(2), 2021, pp. 42–57. DOI: 10.52950/TE.2021.9.2.004.

[9] R. Mauliandri, M. Maimunah, and Y. Roza, Kesesuaian Alat Evaluasi Dengan Indikator Pencapaian Kompetensi Dan Kompetensi Dasar Pada RPP Matematika, J. Cendekia J. Pendidik. Mat., 5(1), 2021, pp. 803–811. DOI: 10.31004/cendekia.v5i1.436.

[10] M. Y. Devi, R. Hidayanthi, and Y. Fitria, Model-Model Evaluasi Pendidikan dan Model Sepuluh Langkah dalam Penilaian, J. Basicedu, 6(1), 2022, pp. 675–683. DOI: 10.31004/basicedu.v6i1.1934.

[11] A. Said and M. Muslimah, Evaluation of Learning Outcomes of Moral Faith Subjects during Covid-19 Pandemic at MIN East Kotawaringin, Bull. Sci. Educ., 1(1), 2021, p. 7. DOI: 10.51278/bse.v1i1.99.

[12] M. Muhammad, H. K. Widyaningrum, A. Al Masjid, K. Komariah, and S. Sumarwati, Pelaksanaan Prosedur Evaluasi Pembelajaran Bahasa Indonesia di SMK Pekanbaru pada Masa Pandemi, Stilistika J. Pendidik. Bhs. dan Sastra, 14(2), 2021. DOI: 10.30651/st.v14i2.8262.

[13] L. W. S. UTAMI, Penggunaan google form dalam evaluasi hasil belajar peserta didik di masa pandemi C0VID-19, Teach. J. Inov. Kegur. dan Ilmu Pendidik., 1(3), 2021, pp. 150–156. DOI: 10.51878/teaching.v1i3.453.

[14] S. Sawaluddin and S. Muhammad, Langkah-Langkah dan Teknik Evaluasi Hasil Belajar Pendidikan Agama Islam, J. PTK dan Pendidik., 6(1), 2020. DOI: 10.18592/ptk.v6i1.3793.

[15] R. Rosnaeni, Karakteristik dan Asesmen Pembelajaran Abad 21, J. Basicedu, 5(5), 2021, pp. 4341–4350. DOI: 10.31004/basicedu.v5i5.1548.

[16] A. Muhammad Santoso and S. Arif, Efektivitas Model Inquiry dengan Pendekatan STEM Education terhadap Kemampuan Berfikir Kritis Peserta Didik, J. Tadris IPA Indones., 1(2), 2021, pp. 73–86. DOI: 10.21154/jtii.v1i2.123.

[17] C. P. Dwyer, M. J. Hogan, and I. Stewart, An integrated critical thinking framework for the 21st

century, Think. Ski. Creat., 12, 2014, pp. 43–52. DOI: 10.1016/j.tsc.2013.12.004.

[18] S. ŽivkoviĿ, A Model of Critical Thinking as an Important Attribute for Success in the 21st Century, Procedia - Soc. Behav. Sci., 232, 2016, pp. 102–108. DOI: 10.1016/j.sbspro.2016.10.034.

[19] A. Fauziana and A. Dessy Wulansari, Analisis Kualitas Butir Soal Ulangan Harian di Sekolah Dasar dengan Model Rasch, Ibriez J. Kependidikan Dasar Islam Berbas. Sains, 6, 2021, pp. 10–19. DOI: 10.21154//ibriez.v6i1.112.

[20] L. N. Oláh, N. R. Lawrence, and M. Riggan, Learning to Learn From Benchmark Assessment Data: How Teachers Analyze Results, Peabody J. Educ., 85(2), 2010, pp. 226–245. DOI: 10.1080/01619561003688688.

[21] "Wikimedia Commons."

[22] R. Mapeala and N. M. Siew, The development and validation of a test of science critical thinking for fifth graders, Springerplus, 4(1), 2015, p. 741. DOI: 10.1186/s40064-015-1535-0.

[23] B. S. & W. Widhiarso, Aplikasi Pemodelan Rasch Pada Assessment Pendidikan, Trim Komunikata, 2015.

[24] V. A. Scholtes, C. B. Terwee, and R. W. Poolman, What makes a measurement instrument valid and reliable?, Injury, 42(3), 2011, pp. 236–240. DOI: 10.1016/j.injury.2010.11.042.

[25] D. W. Russell, UCLA Loneliness Scale (Version 3): Reliability, Validity, and Factor Structure, J. Pers. Assess., 66(1), 1996, pp. 20–40. DOI: 10.1207/s15327752jpa6601_2.

[26] S. H. Carson, J. B. Peterson, and D. M. Higgins, Reliability, Validity, and Factor Structure of the Creative Achievement Questionnaire, Creat. Res. J., 17(1), 2005, pp. 37–50. DOI: 10.1207/s15326934crj1701_4.

[27] S. Azwar, Reliabilitas, Validitas, Interpretasi dan Komputasi, Yogyakarta: Liberty, 2006.

[28] C. D. Hopkins and R. L. Antes, Classroom Measurement and Evaluation, Illionis: F.E. Peacock, 1999.

[29] S. Mui Lim, S. Rodger, and T. Brown, Using Rasch analysis to establish the construct validity of rehabilitation assessment tools, Int. J. Ther. Rehabil., 16(5), 2009, pp. 251–260. DOI: 10.12968/ijtr.2009.16.5.42102.

[30] F. Donkor, The comparative instructional effectiveness of print-based and video-based instructional materials for teaching practical skills at a distance, Int. Rev. Res. Open Distrib. Learn., 11(1), 2010, p. 96. DOI: 10.19173/irrodl.v11i1.792.