

## **PORTRAIT OF THE QUALITY OF *MULTIPLE-CHOICE TESTS* FOR ENGLISH LANGUAGE SUBJECTS ON STUDENTS**

Wawan Gunawan

Postgraduate School of Educational Technology, PGRI Adi Buana University

[wawan.gunawan@unipasby.ac.id](mailto:wawan.gunawan@unipasby.ac.id)

### **ABSTRACT**

*This study aims to analyze test items on students' understanding in learning English. Test items used as an assessment must have objective quality that can be accounted for, so the questions are analyzed statistically which includes (1) validity analysis, (2) reliability analysis, (3) difficulty level analysis, and (4) question discrimination analysis. The form of research used in this study is quantitative descriptive research. This research is an evaluation involving 30 students. The instrument used in this study is data collection through documentation techniques. Analysis of the test item data was carried out using the SPSS 25 program. The results of the difficulty index test show that the test items are in the very easy category 20%, the test items are in the medium category 72% and the test items are in the easy category 20%. While the test results of the test show that the test items are at a very good level with a presentation of 4%, the questions are in the good category with a presentation of 64%, and the questions are in the sufficient category with a presentation of 24%. The questions presented can be analyzed with valid and reliable results and can be distributed well. Thus, it can be concluded that the questions are suitable for use in midterm exams. The benefits of this research include the need to always prepare backup questions to replace those deemed unsatisfactory, revisions to questions deemed good enough for reuse, and addition of questions deemed good to the question bank.*

**Keywords:** *Item Analysis, Multiple Choice, English Subject*

### **A. Introduction**

Two important components in the world of education include teachers and students. Teachers who guide students in learning and students who are the objects in the learning (Inawati et al., 2022; Putri Andini & Mukhlis, 2023; Verawati et al., 2023) . Pradani & Efendi, (2023) said that teacher competency standards include 4 main competencies, namely pedagogical

competence, personality competence, social competence, and professional competence. Pedagogical competence is the ability of teachers in managing student learning which at least includes the following: (a) understanding the insight or foundation of education, (b) understanding students, (c) curriculum/syllabus development, (d) learning planning, (e) implementation of educational and dialogical learning,

(f) utilization of learning technology; g) evaluation of learning outcomes (EHB), and (h) development of students to actualize their various potentials.

In the world of education, evaluation is conducted to determine the quality of the educational components. For example, a teacher conducts an evaluation of students, the goal is to determine the extent of the student's abilities. Evaluation or assessment is an umbrella concept that covers several concepts under it, such as measurement and testing. Evaluation refers to or contains the meaning: an action or a process to determine the value of something. Assessment is the process of giving or determining a value to a particular object based on certain criteria. Assessment is carried out with the aim of assessing the process and results of student learning in school, diagnosing student learning difficulties and determining class promotion or graduation. The function of assessment is as a tool to determine whether instructional objectives have been achieved, provide feedback for improving the teaching and learning process, increase student learning motivation, and serve as a basis for

compiling student learning progress reports to parents (Hidayatunnisa'i et al., 2023; Nur Cahyo et al., 2022; Prawiki & Helendra, 2022) .

Broadly speaking, the evaluation tools used can be classified into two types, namely tests and non-tests. Test techniques can be carried out both written and unwritten. Non-test techniques are used to assess attitudes, skills, behavior, and so on. Tests in education are assessment tools or assessment methods that are systematic, valid, reliable and objective to determine students' abilities, skills, and level of knowledge of teaching materials, in the form of a task or problem that must be completed by a student or group of students (Hanan et al., 2023; Kusumayanti & Jannah, 2022; Rismaulhijjah & Kuswanti, 2022) . Based on the objectivity of scoring, tests are divided into objective tests and subjective tests. Multiple-choice tests are a form of text that has a correct or most appropriate answer. In this test, the correctness of the answer is absolute. Answers only have four possibilities: A, B, C, and D.

English subject tests, in relation to or as part of English language learning, are tools used to assess

students' mastery of the English they have learned. English subject tests are a tool for measuring student ability. Item analysis is an activity teachers must undertake to improve the quality of the questions they have prepared. It aims to identify good, bad, and poor questions. Question analysis can provide information about a question's weaknesses and provide "guidance" for improvement.

In this case, the author will discuss the study of test content, namely analyzing the level of validity, reliability, level of difficulty, and discrimination. Mira et al., (2022); RP Sari, (2022); VNI Sari et al., (2022) , As the name suggests, a teacher-made test is a test created by the teacher themselves. The test is intended to measure the level of student success in achieving objectives after the teaching process managed by the teacher concerned.

## **B. Research Methods**

The research method used in this study is a descriptive research method through a quantitative approach. According to (Sugiyono, 2017, 2019), the quantitative descriptive research method aims to describe a phenomenon, event,

symptom, and incident that occurs factually, systematically, and accurately. Phenomena can be in the form of forms, activities, relationships, characteristics, and similarities and differences between phenomena. The study was conducted in the odd semester of the 2024/2025 academic year using multiple-choice questions designed to determine students' understanding of the English subject.

**Table 1. English subjects**

No	Learning objectives
1	Identifying parts of English and describing their functions
2	Describe the English language process and relate the importance of this process to living things.
3	Identify and create simulations using charts or simple tools about the English cycle.

Measurement of the level of difficulty of the test items is generated based on a comparison of the number of incorrect and correct responses (Jumini et al., 2023; Marambaawang et al., 2023; Riau & Mukhlis, 2023) which are then compared with the test item difficulty index criteria in Table 2.

**Table 2. Criteria for the item difficulty index**

Difficulty Index Value	Interpretation
0.00-0.15	very difficult
0.16-0.30	difficult
0.31-0.70	currently
0.71-0.85	easy
0.86-1.00	very easy

The item discrimination index is classified based on the quality of the item, which can be referenced from the final score obtained by each respondent (Hanan et al., 2023). This value is used as a reference in the Pearson product moment test to classify the calculated  $r$  value for each item to obtain the discrimination index. There are six categories of discrimination index classifications, as shown in Table 3.

**Table 3. Category of item discrimination index**

Distinguishing Power Index Value	Criteria
<b>Negative</b> <b>0 - 0.190.20 - 0.290.30 - 0.39= 0.4 &gt; 0.4</b>	Damaged Bad Average Good Very Good Excellent

Source: (Alfiana et al., 2021)

### C. Research Result

Validity testing is carried out in two ways, namely (1) comparing the Pearson correlation value or calculated  $r$  value with the  $r$  table value, the calculated  $r$  value must be greater than the  $r$  table value ( $r_{\text{calculated}} > r_{\text{table}}$ ). 2) Combining the Pearson correlation value with the significance obtained, the value obtained must be positive and supported by a significance value  $< 0.05$ . The test results using Pearson correlation can be seen in table 4.

**Table 4. Validity test results**

No Question	R Calculation Results	Significant Value	Validity
item01	.631 **	0.000	valid
item 02	0.318	0.087	invalid
item 03	.635 **	0.000	valid
item 04	.696 **	0.000	valid
item05	.486 **	0.007	valid
item 06	.512 **	0.004	valid
item 07	.528 **	0.003	valid
item 08	.488 **	0.006	valid
item 09	.474 **	0.008	valid
item 10	.440 *	0.015	valid
item 11	0.231	0.220	invalid
item 12	.671 **	0.000	valid
item 13	.393 *	0.032	valid
item 14	.789 **	0.000	valid
item 15	.538 **	0.002	valid
item 16	.465 **	0.010	valid
item 17	.482 **	0.007	valid
item 18	.520 **	0.003	valid
item 19	.696 **	0.000	valid
item 20	.577 **	0.001	valid
item 21	.465 **	0.010	valid
item 22	.482 **	0.007	valid
item 23	.520 **	0.003	valid
item 24	.696 **	0.000	valid
item 25	.577 **	0.001	valid

In this study, based on the data above, it can be stated that the validity test of the questions is categorized as valid with a high value of 789 and the highest significant value is 0.000 so that the instrument is suitable for use as a measuring tool in evaluating English language learning.

The reliability test based on Chronbach's alpha compares the test result value with a value of 0.06. If the test result value is greater than 0.06, the item is considered reliable. The data are displayed in Table 5.

**Table 5. Reliability Test Reliability Statistics**

Cronbach's Alpha	N of Items
.901	23

The feasibility parameters of an instrument are not only determined by validity and reliability but must also be able to describe the differences in knowledge and abilities of students. Therefore, an instrument of difficulty index and question discrimination power is needed to be able to differentiate the ability level of each student (Ismiyati et al., 2023; Rahayu et al., 2023; RP Sari, 2022).

The item difficulty index was conducted to identify the level of difficulty of each item by comparing the test results and the item difficulty index in Table 2. The results of the difficulty index test showed that 20% of the items were in the very easy category, 72% of the items were in the medium category, and 20% of the items were in the easy category. More detailed results are presented in Table 6.

**Table 6. Item difficulty index**

No Question	R-value calculation	Decision Making Criteria	Criteria
1	0.60	Difficulty level index 0.00-0.15 = very difficult 0.16-0.30 = difficult 0.31-0.70 = moderate 0.71-0.85 = easy 0.86-1.00 = very easy	Currently
2			
3	0.66		Currently
4	0.50		Currently
5	0.70		Currently
6	0.60		Currently
7	0.76		Easy
8	0.66		Currently
9	0.73		Easy
10	0.76		Easy
11			
12	0.53		Currently
13	0.70		Currently
14	0.53		Currently
15	0.63		Currently
16	0.76		Easy
17	0.63		Currently
18	0.70		Currently
19	0.60		currently
20	0.60		currently
21	0.76		Easy
22	0.63		Currently
23	0.70		Currently
24	0.60		Currently
25	0.60		Currently

The item discrimination test serves to illustrate the differences in ability among respondents and compare them to the item discrimination category. The test results indicate that 4% of the items were in the excellent category, 64% were in the good category, and 24% were in the fair category. More detailed results are presented in Table 7.

**Table 7. Results of the test of the discriminating power of the questions**

No Question	R Calculation Results	Interpretation criteria for differential power	Validity
1	0.576		Good
2	0.593		Good
3	0.593		Good
4	0.659		Good
5	0.471		Good
6	0.454		Good
7	0.481		Good
8	0.403		Good
9	0.387		Enough
10	0.411		Good
11			
12	0.612		Good
13	0.292		Enough
14	0.761		very well
15	0.489		Good
16	0.398		Enough
17	0.378		Enough
18	0.510		Good
19	0.616		Good
20	0.515		Good
21	0.398		Enough
22	0.378		Enough
23	0.510		Good
24	0.676		Good
25	0.515		Good

In this study, the results of the test on the discriminating power of the questions indicated that all the questions were in the sufficient and good categories. Therefore, it can be concluded that all the questions were of good enough quality that they did not require improvement and could be used as an evaluation tool in learning.

#### D. Conclusion

The development of an evaluation measuring tool in the form of multiple-choice questions to measure conceptual understanding of English language learning has been successful. The analysis results show that all questions tested are valid and reliable, so they can be used as a quality evaluation tool. The results of the difficulty index test show that 20% of the questions are in the very easy category, 72% are in the moderate category, and 20% are in the easy

category. Meanwhile, the test of the question discrimination power shows that 4% of the questions are at the very good level, 64% are in the good category, and 24% are in the sufficient category. The questions presented can be analyzed with valid and reliable results and can be distributed well. Therefore, it can be concluded that the questions are suitable for use in midterm exams. The benefits that can be obtained from the results of this study are that it is advisable to always prepare reserve questions to replace questions in the poor category, revisions can be made to questions in the good enough category for subsequent reuse, and questions in the good category can be entered into the question bank.

#### BIBLIOGRAPHY

- Hanan, MP, Jannah, RR, & Alim, JA (2023). Analysis of Mathematics Question Items on Story-Based KPK and GPB Material at Sdn 111 Pekanbaru. *Journal of Educational Learning and Innovation (Elia)*, 3 (1). <https://doi.org/10.46229/Elia.V3i1.538>
- Hidayatunnisa'i, Kusriani, & Kusnawi. (2023). Comparison of the Performance of the Naive Bayes Method and Support Machine

- (SVM) in Analyzing the Quality of Multiple-Choice Questions. *Jurnal Fasilkom*, 13 (02). <https://doi.org/10.37859/Jf.V13i02.5087>
- Inawati, Arya, A., & Lasmiatun. (2022). Analysis of the Difficulty Level of Teacher-Made Indonesian Language Subject Questions. *Baistrando*, Vol 1 (No 2).
- Ismiyati, I., Raharjo, TH, Tusyanah, T., & Sholikah, M. (2023). Item Analysis Training Based on Classical Test Theory Assisted by IteMan to Improve the Quality of Assessment Instruments. *Japi (Journal of Indonesian Community Service Access)*, 8 (2). <https://doi.org/10.33366/Japi.V8i2.5064>
- Jumini, S., Madnasri, S., Cahyono, E., & Parmin, P. (2023). Analysis of the Quality of Scientific Literacy Measurement Items Using Classical Test Theory and the Rasch Model. *Proceedings of the National Seminar on Postgraduate Studies at Unnes*.
- Kusumayanti, A., & Jannah, N. (2022). Analysis of Independent Entrance Examination Question Items at Uin Alauddin Makassar Using Modern Test Theory. *Mapan: Journal of Mathematics and Learning*, 10 (1).
- Marambaawang, D., Bano, VO, & Hada Enda, RR (2023). Analysis of the Quality of Final Assessment Questions for the Odd Semester of 2021/2022 Using IteMan at Smp Negeri 1 Kambera. *Dharmas Education Journal (De\_Journal)*, 4 (1). <https://doi.org/10.56667/Dejournal.V4i1.961>
- Mira, K., Hayon, VHB, & Tinenti, YR (2022). Analysis of Multiple Choice Questions on the Topic of Electrolyte and Non-electrolyte Solutions for Class X Science at Sma Negeri 1 Lamba Leda, East Manggarai. *Wahana*, 73 (2). <https://doi.org/10.36456/Wahana.V73i2.4201>
- Nur Cahyo, A., Luriawati, D., & Wagiran, W. (2022). Analysis of Language Skills Assessment Items in Explanatory Text Learning for Grade XI. *Jbsi: Indonesian Language and Literature Journal*, 2 (01). <https://doi.org/10.47709/Jbsi.V2i01.1493>
- Pradani, RA, & Efendi, A. (2023). Analysis of School Exam Questions Using the IteMan Program. *Indonesian Language Education and Literature*, 8 (2). <https://doi.org/10.24235/ileal.V8i2.11002>
- Prawiki, SM, & Helendra. (2022). Analysis of the Quality of Final Exam Questions for Odd Semester 2020/2021 Academic Year Biology Subject for Class X of Sma Negeri 1 Teluk Sebong. *Biodidactics: Journal of Biology and Its Learning*, 17 (2).

- Putri Andini, D., & Mukhlis, M. (2023). Analysis of Question Items in the Minimum Competency Assessment Instrument for Reading Literacy at SMP IT Insan Utama Pekanbaru. *Diglossia: Journal of Language, Literature, and Teaching Studies*, 6 (2). <https://Doi.Org/10.30872/Diglossia.V6i2.658>
- Rahayu, W., Sulaeman, E., Arnawisuda Ningsi, B., Arofah, I., & Akbari, W. (2023). Analysis of Elementary School Teachers' Comprehension Ability in Developing Numerical Literacy-Based Test Items. *Perduli: Journal of Community Service*, 4 (01). <https://Doi.Org/10.21009/Perduli.V4i01.30490>
- Riau, UI, & Mukhlis, M. (2023). Analysis of Reading Literacy Items in the Minimum Competency Assessment Instrument for Vocational High School Students in Indonesian and Regional Literature (Vol. 13, Issue 2).
- Rismaulhijjah, W., & Kuswanti, N. (2022). Analysis of Daily Test Items Resulting from Teacher Development of Human Movement System Material for Grade XI Science. *Biology Education Scientific Periodical (Bioedu)*, 11 (3). <https://Doi.Org/10.26740/Bioedu.V11n3.P643-661>
- Sari, RP (2022). Analysis of Final Exam Questions for Even Semester Mathematics Subject for Grade XI of State Vocational High School 1 Percut Sei Tuan Academic Year 2019/2020. *Journal of Science Education Research*, 7 (1). <https://Doi.Org/10.32696/Jp2mipa.V7i1.1360>
- Sari, VNI, Utomo, APY, & Sumarwati. (2022). The Quality of Indonesian Language Questions at Muhammadiyah 1 Middle School, Pontianak: Analysis of Vina Question Items. *Journal of Indonesian Language and Literature Education*, 11 (2).
- Sugiyono. (2017). *Qualitative Quantitative Research Methods and R&D (25th Ed.)*. Alfabeta.
- Sugiyono, S. (2019). *Quantitative, Qualitative, and R&D Research Methods (Sutopo (Ed.); 1st Ed.)*. Alfabeta.
- Verawati, Y., Siskawati, FS, & Susilaningtyas, T. (2023). Analysis of Final Semester Exam (Uas) Question Items for Mathematics Subject in the 2020/2021 Academic Year for Grade VII of At Tanwir Islamic Junior High School, Ledokombo District, Jember Regency. *Jurnal Jendela Pendidikan*, 3 (01). <https://Doi.Org/10.57008/Jjp.V3i01.422>