Bulletin of Science Education

Vol. 4, No. 3, December 2024
https://attractivejournal.com/index.php/bse/index

An Application of Difficulty Level Analysis of Question Items in Language Learning Evaluation

Ni Putu Eka Maryani Dewi[1*], Ida Bagus Putrayasa[1]
[1] *Universitas Pendidikan Ganesha, Indonesia*

ekameyani88@gmail.com*

ARTICLE INFO
Article history:
Received
October, 2024
Revised
December 02, 2024
Accepted
December 23, 2024

Abstract

One important component of learning is the assessment or evaluation process. Most evaluations will use items as tests, but not necessarily all items used are of high quality due to unvalidated difficulty levels. This article aims to provide an understanding of item difficulty and how to measure it as well as the importance of analysing item difficulty in language learning tests. The method used in this article is a literature review that examines the application of analysing the difficulty level of test items in language learning and its importance for learning. The results obtained are that it is important to analyse the level of difficulty of the questions to determine the quality of the test so that it can be feedback for learning; increase the validity and reliability of the test, adjust students' abilities, and improve the quality of the test for the next time. The formula used in item analysis includes a difficulty index denoted by the letter P (proportion), with an interval of 0.30-0.70 said to be a good question (enough).

Keywords: Question Item Analysis, Difficulty Level, Learning Evaluation

Published by CV. Creative Tugu Pena
ISSN 2774-4299
Website https://attractivejournal.com/index.php/bse/

PENDAHULUAN

Learning evaluation is an important component in the educational process. Assessment or evaluation is one of the tasks of a teacher in an effort to find out the extent to which the learning objectives that have been set have been achieved. Law no. 14 of 2005 article 10 paragraph 1 that teachers as determinants of the success of the teaching and learning process must be competent in designing, implementing, evaluating and revising the questions used in evaluation activities which determine the success or failure of the learning process that has been carried out. In line with this, assessment functions as a tool to determine the success of student learning processes and outcomes (Arif, 2014). Assessment is generally divided into three domains, namely the cognitive domain, affective domain and psychomotor domain. The domain most often assessed by teachers in schools is the cognitive domain because it is directly related to students' understanding of mastering the content of teaching material.

There are several techniques that teachers can choose to carry out evaluations. There are two types of evaluation techniques, namely test techniques and non-test techniques. Test techniques can be carried out in written or unwritten form. Meanwhile, non-test techniques are usually used to assess students' attitudes, behavior and personality during teaching and learning activities in class Permatasari, 2014). There are 2

types of instruments in learning evaluation, namely objective tests and non-objective tests. The objective test is divided into 4 which include: multiple choice questions, true and false choices, matching and short answers. Meanwhile, non-objective tests take the form of long descriptions (Wulan, 2018).

However, often the evaluation results are not as expected. A crucial aspect in evaluation that may be the cause is the quality of the assessment tools, one of which is the question items used. Good test items should not only be able to accurately measure student learning achievement, but also provide valuable feedback for teachers to improve the learning process. Question item analysis is a systematic effort to assess the quality of each question item used in an evaluation.

Item analysis is an important step in evaluating language learning. Through this analysis, you can determine the quality of each question, identify areas that need improvement, and increase learning effectiveness. In the language context, item analysis is carried out to assess how well the questions are able to measure language skills such as listening, speaking, reading and writing. The main purpose of analyzing items in a test made by a teacher is to identify deficiencies in the test or in learning (Anastasi and Urbina, 1997: 184). Mustafa (2012) also provides an explanation of the function of item analysis in more detail, namely to 1) find out questions that are difficult for test takers to do, so that the teacher will guide them specifically on material that is considered difficult; 2) analyze the different power levels of the questions so that it can be used as a basis for the teacher in determining which questions need to be maintained and which need to be revised.

Based on this objective, the item analysis activity has many benefits, including, (1) it can help test users in evaluating the tests used, (2) it is very relevant for the preparation of informal and local tests prepared by teachers for students in class, (3) supports the writing of effective questions, (4) can materially improve tests in class, (5) increases question validity and reliability (Anastasi and Urbina, 1997: 172).

Analysis of question items is an important stage in learning evaluation to ensure the questions are of good quality. Question analysis involves aspects, one of which is analysis of the difficulty level of the question items. Questions with a difficulty level that is too low or high will not be able to provide optimal information about students' skills. Questions that are too easy do not provide a challenge for students with high abilities, while questions that are too difficult can prevent students from demonstrating knowledge or skills effectively.

This article will discuss the meaning of the level of difficulty of test items, how to measure or analyze the level of difficulty of test items, as well as the importance of item analysis in evaluating language learning.

METHOD

The method used in this article is a literature review or literature review. This method collects several previous studies to examine the application of test item difficulty level analysis in language learning and its importance for learning. Zed in Kartiningsih's research (2015) said that the literature study method is a series of activities related to methods of collecting library data, reading and taking notes, and managing research materials.

DISCUSSION

Level of Difficulty of Question Details in Language Learning

The difficulty level of a question is the opportunity to answer a question correctly at a certain level of ability or it could be said to know whether a question is easy or difficult (Fatimah, Alfath, 2019). Whether or not the learning outcomes test items are of good quality can first be determined from the degree of difficulty or level of difficulty of each item. Item items can be said to be good if the items are neither too difficult nor too

easy. So, if all testees cannot answer the question correctly, (because it is too difficult) it cannot be called a good item. Even if all testees can answer correctly, (because the questions are too easy) they cannot be included in the good category.

According to Azis (2016) the level of difficulty is how easy or difficult an item is for a group of students. In general, it can be said that the level of difficulty is the level of whether a problem is easy or not given to a group of students. The level of difficulty of a question is the proportion of students who answer a question correctly. The level of difficulty of the items means reviewing the test questions in terms of their difficulty with the aim of getting information about which questions are classified as easy, medium and difficult (Valen & Satria, 2021).

The difficulty level of a question item refers to the level of difficulty or challenge provided by a question in an exam or test. The difficulty level of questions is used to categorize questions based on the level of complexity or level of understanding required by students to answer them. This category is very important in preparing tests or exams, because it can help examiners to assess the extent to which participants can master the material that has been taught. The difficulty level of the questions is used to categorize the questions, with 3 categories, namely easy, medium and difficult (Dewi, et al, 2018). Each category can be interpreted as: (1) **Easy**: Easy questions can usually be answered quickly and accurately by most students. These questions generally test basic understanding, facts, or concepts that have been taught directly. Easy questions have a high success rate, so the majority of students can answer them correctly. (2) **Currently**: Questions with a moderate level of difficulty test students' deeper understanding or ability to apply the concepts or knowledge they have learned. These questions usually contain little challenge or require simple analysis. Most students can answer it correctly, but there may be difficulties, depending on their understanding and abilities. (3) **Difficult**: Difficult questions usually test higher-order thinking skills, such as analysis, synthesis, or evaluation. These questions often contain questions that require very in-depth understanding or application of knowledge in more complex situations. Only a small percentage of students can answer the questions correctly, because these questions are more challenging and require more time and effort to complete.

How to Measure the Difficulty Level of Question Details

Fatimah and Alfath, 2019 quoted from a book entitled Psychological Education, Witherington which said that whether or not the degree of difficulty of a learning outcome test item is adequate can be determined from the size of the number that represents the level of difficulty of the item. Numbers that can give an indication of the difficulty level of the item are known as terms *difficulty index* (item difficulty index number), which in the world of learning outcomes evaluation is generally denoted by the letter P, namely *Proportion*. A test should not be too easy, nor should it be too difficult.

According to Fitriani (2021), the level of difficulty of a question can be seen from the size of the difficulty index. The difficulty level of a question is the opportunity to answer a question correctly at a certain level of ability which is usually expressed in the form of an index. The results of calculating the difficulty index are interpreted in three criteria, namely: P= 0.00-0.30 is a difficult question, P= 0.31-0.70 is a question with a medium level of difficulty, and P= 0.71-1, 00 is an easy matter.

The difficulty index is expressed in the form of a proportion between 0.00-1.00. The smaller the difficulty index number, the more difficult the question is. Test items can be declared as good items if they are neither too difficult nor too easy. In other words, the level of difficulty of the items is moderate or sufficient, that is, they have a difficulty index between 0.31-0.70.

The questions used can be made in multiple choice or essay form. According to Iskandar (2015), multiple choice questions can be made into several types of format,

including: (1) conventional, where the question has several answer choices but only contains one true statement. (2) *alternative choice*, namely questions that only provide two types of answer choices; right answer and wrong answer. (3) *matching*, this type of question simply provides a series of choices that can be applied to two or more items. (4) *true-false*, this format is in the form of a statement that will be answered by students by responding to the contents of the statement with a right or wrong answer. (5) *pictorial item set*, this question format is designed based on features such as graphs, tables, photos, works of art, or other visual forms followed by a series of questions that correlate with each other. (6) *problem-solving item set*, displays an image in the form of a description or features such as format. Multiple choice questions are used to determine student learning outcomes in terms of knowledge, namely memory, understanding, application, analysis, synthesis and evaluation. The advantage of this type of question is that it can measure good memory and high mental processes; assessed quickly and efficiently and easy to measure the level of difficulty (Saputra, et al, 2022).

The way to analyze multiple choice questions was proposed by Arikunto (2013) which is in line with his opinion above, stating that a question is considered good if it has a medium level of difficulty, namely between 0.31-0.70.

The item difficulty index number can be obtained using the following formula:

$$P = \frac{B}{JS}$$

P is the difficulty index number, B = the number of testees who can answer correctly the item in question, and JS = the number of testees who take the learning outcomes test.

Another formula commonly used to measure the level of difficulty of a question item is:

$$P = \frac{Np}{N}$$

Where P is the proportion or item difficulty index number. Np is the number of testees who answered the item correctly. N is the number of testees who took the learning outcomes test.

Criteria The difficulty index of questions is interpreted by Robert L. Thorndike and Elizabeth Hagen as follows:

Table. 1

The magnitude of P	Interpretation
Less than 0.30	Too hard
0,3 - 0,70	Fair (medium)
More than 0.70	Too easy

Meanwhile, according to Whiterington in his book Psychological Education as follows:

Table. 2

The magnitude of P	Interpretation
Less than 0.25	Too hard
0,25 - 0,75	Fair (medium)
More than 0.75	Too easy

Meanwhile, to analyze the level of difficulty of essay questions, you can use the following formula:

$$Mean = \frac{Jumlah\ skor\ siswa\ pada\ butir\ soal\ tertentu}{Banyak\ siswa\ yang\ mengikuti\ tes}$$

It can be concluded that the item difficulty index criteria can use the following intervals:

Table 3. Question Difficulty Index Criteria

Interval	Interpretation
0,00 - 0,30	Difficult category questions
0,31 - 0,70	Medium category questions
0,71 - 1,00	Easy category questions

When preparing question papers for exams or tests, the difficulty level of the questions should be balanced, calculating 25% of the questions in the difficult category, 50% of the questions in the medium category, and 25% of the questions in the easy category (Fatimah, Alfath, 2019).

The Importance of Analyzing the Level of Difficulty of Questions in Tests

The importance of teachers knowing the difficulty level of questions is to minimize tests that are too easy and too difficult. As Arikunto (1999: 210) said, determining the appropriateness of questions requires paying attention to the purpose of using the questions. If test questions are used to achieve student learning outcomes, then the test questions tend to use questions that are not too easy and not too difficult. By analyzing the difficulty level of the questions after the test is carried out, you will be able to find out what percentage of the questions are difficult, medium and easy and will help teachers to know which questions need to be revised and which need to be maintained.

It is not only the student's competence that needs to be evaluated, but the questions given to students also need to be analyzed to determine the effectiveness and quality of a question item. Therefore, it is important to analyze the questions to find out the quality of a question item (Ratni, 2023). The difficulty level of the questions has 2 uses, namely for teachers and for testing and teaching. For teachers, it is an introduction to the concept of re-learning and providing input to students about their learning outcomes, obtaining information about curriculum emphasis or suspecting biased test items. Its uses for testing and teaching are, as an introduction to concepts that need to be retaught, signs of the strengths and weaknesses of the school curriculum, providing input to students, signs of possible biased questions and assembling tests that have accurate question data.

Item analysis is important as a means of evaluating the use of multiple choice questions. The results of the analysis can be used as a means of evaluation to improve the questions so that they become better. Improved questions will be more effective for use in determining the quality of learning that has been implemented (Wiyasa et al., 2019). On the other hand, the analysis carried out will provide specific information regarding the question item. This specific information and analysis findings are an effort to improve learning and improve the quality of education so that educators are not careless in conducting evaluations because they are trying to measure it using several categories. In analyzing, we still prioritize the specified standards (Metsämuuronen in Sari, et al, 2022).

Sudijono (2012) revealed several follow-up actions that can be carried out after analyzing the level of difficulty of the questions, as follows:

a. Questions that have a level of difficulty in the good category (medium level of difficulty) should be saved in the question bank so that they can be used again in the future.

b. For items that are in the difficult category, there are three possible follow-up actions, namely: 1) These items are discarded and will not be issued again in the learning outcomes test in the future. 2) The question items are re-examined to identify the factors that cause students to have difficulty answering them. Improvements can be made by changing the sentence so as not to cause misinterpretation or changing the numbers/nominals in the calculation questions. After corrections have been made, the questions can be saved to the question bank and issued again at a future time. 3) These question items are retained to be used again in tests that are very strict, meaning that most test takers will not pass the selection test.

c. For questions that are included in the easy category, there are three possible follow-up actions, namely: 1) These question items are discarded and will not be issued again in future learning outcomes tests. 2) The question items were re-examined to find out the factors that caused almost all students taking the test to answer correctly. There is a possibility that the alternatives attached to the questions are too easy for test takers to guess. Improvements can be made by improving the answer options or making the question sentences more complex. After corrections have been made, the questions can be saved into the question bank and issued in future learning outcomes tests. 3) Question items are maintained and utilized in loose tests, meaning that the majority of test takers will be declared to have passed the selection. Under these conditions, the test is just a formality.

In line with Sudijono, Susanto, et al. (2015) Question items included in the medium category should be recorded in the question bank book. Furthermore, these question items can be issued again in the learning outcomes test in the future. For difficult question items, there may be three follow-up actions, namely: (1) the question item is no longer used and is not removed again in the next test results. (2) reviewed, researched and explored so that the factors causing the child cannot answer the question can be identified. (3) Difficult questions should have benefits, namely that they can be used in selection tests which are loose, in the sense that most of the testees will be declared to have passed the selection test. In this condition, giving easy questions will give many testees the opportunity to pass the selection test or exam being held.

Research conducted by Magdalena, et al (2021) with the title 'Analysis of Validity, Reliability, Level of Difficulty and Differential Power of Final Exam Question Items for Theme 7 Class III SDN Karet 1 Sepatan' stated the results that of the 20 questions created there were 16 questions (80%) in the medium category, 4 questions (20%) in the difficult category. It can be seen that the questions made in the medium category have exceeded the difficult category. It can be interpreted that the multiple choice questions have good quality, apart from fulfilling validity and reliability, there is a balance in the level of difficulty of the questions. The comparison for making the difficulty level of questions can be made 3-4-3 or 3-5-2 which means 30% easy questions, 40% medium questions, and 30% difficult questions or 30% easy questions, 50% medium questions, and 20% difficult questions. difficult category.

Another similar research is research conducted by Son (2019) with the title 'Mathematical Problem Solving Ability Instrument' which analyzes problem solving questions in the form of descriptions because the problem solving steps are simpler so it is easy to analyze student answers. The questions analyzed were four items related to the Algebraic Operation Lesson material. Of the four description questions, there is 1 question in the difficult category, 2 questions in the medium category, and 1 question in the difficult category. This shows that the proportionality of the four questions is normal or sufficient, namely 25% in the easy category, 50% in the medium category, and 25% in the difficult category.

The third relevant research was written by Himawan, et al (2022) with the title 'Analysis of Odd Semester Final Assessment Practice Questions for Indonesian Language

Subjects' using the ITEMAN computer program by analyzing 40 multiple choice questions. The results obtained were 6 questions in the very easy category, 17 questions in the easy category, 8 questions in the medium category, and 9 questions in the difficult category. The results of this analysis did not achieve a balance of items to be used as a test because the percentage of easy and difficult items was greater than the items in the medium category.

The latest research designed by Forniawan, et al (2023) with the research title 'Follow-up Analysis of Cognitive Diagnostic Assessment Results for Natural Sciences Subjects' found that of the 10 questions analyzed, 60% of the questions had a difficulty level in the easy category and 40% in the medium category without any questions in the difficult category. This shows that the questions used are not too challenging for students. The results of the analysis indicate that there is a need to improve some of the questions so that they can be prepared with a more varied and balanced level of difficulty.

From several studies above, as a question maker, you can follow up on questions that are in the very easy and difficult categories to be reviewed, or are not used again in the next test. In this way, the quality of the questions will be better and more valid for assessing student abilities. Teachers also receive feedback to make improvements in implementing differentiated learning according to students' level of understanding. This will make the foundation of students' basic knowledge and skills stronger before learning other knowledge and skills that have a higher level of difficulty (Dewi et al, 2023). Data from the results of working on assessment items can be used to identify the level of difficulty of assessment items and their suitability and patterns of student ability categories and their suitability. These results can then provide reflections and recommendations to teachers as learning implementers to improve the learning process that has been carried out (Rahman, et al, 2022).

CONCLUSION

Analysis of the level of difficulty of the questions is an important step that must be implemented in evaluating a test or exam. The main purpose of this analysis is to find out how easy or difficult a question is for test takers. The questions can be categorized into three, namely questions with easy, medium and easy levels of difficulty. The ideal questions to use are questions with a sufficient or medium level of difficulty. However, the proportion must also be balanced by adding questions with high and low levels of difficulty (difficult questions and easy questions). Analysis of test items can be carried out using the difficulty index formula, denoted as P or proportion. It is important to apply question item analysis to determine the quality and effectiveness of the questions. This application is needed for teachers and teaching so that it can provide feedback and improve questions on the next test. Follow-up actions on the three categories of questions can also be implemented, not just by removing them from the test.

REFERENCES

Arif, Muchamad. (2014). Implementation of the Anates Application in the form of multiple choice questions. *Edutic Scientific Journal*, Vol. 1, No. 1. ISSN 2407-4489.

Arikunto. Suharsimi. (2018). Basics of Educational Evaluation. In Jakarta: Bumi Aksara (p. 46).

Azis. (2016). Analysis of Teacher-Made Tests Field of Study Mathematics Class V SD 1 Katobengke. ISSN:2008-2157 Vol. 6 No. 1, 18.

Dewi, N. L., Sukamto, & Prasetyowati, D. (2023). Analysis of Diagnostic Assessment Results in Natural and Social Science Subjects in Grade IV Elementary School. *Didactics: PGSD STKIP Subang Scientific Journal*, Vol. 9, No. 02, 4979–499.

Dewi, S.S., Hariastuti, R.M., & Utami, A.U. (2019). Analysis of the Level of Difficulty and Differentiating Power of the 2018 Middle School Mathematics Olympiad (OMI)

Questions. *Transformation: Journal of Mathematics and Mathematics education*. Vol. 3, *No.* 1, 15-26. https://doi.org/10.36526/tr.v3i1.388

Fahlevi, M. S., Zanthy, L. S. (2020). Analysis of Students' Difficulties in Completing Description Questions on Flat Side Building Material. *Journal of Innovative Mathematics Learning*, Vol. 3, No. 4. E-ISSN: 2614-2155.

Fatimah, L.U., & Alfath, K. (2019). Analysis of Question Difficulty, Discriminatory Power, and Distractor Functions. *Journal of Islamic Communication and Education*, Vol. 8, No. 2. TWO: https://doi.org/10.36668/jal.v8i2.115

Fitriani, Nani. (2021). Analysis of Level of Difficulty, Differentiating Power, and Effectiveness of Distracters on Maternal and Neonatal Emergency Awareness Training Questions. *Paedagoria: Journal of Educational Studies, Research and Development*, Vol. 12, No. 2. ISSN 2614-3674.

Formiawan, A., Wati, D. R. (2023). Follow-up Analysis of Cognitive Diagnostic Assessment Results for Natural Science Subjects. *Al-Jahiz: Journal of Biology Education Research*, Vol. 4, No. 2. E-ISSN: 2722-5275.

Himawan, R., Nurgiyantoro, B. (2022). Analysis of Final Odd Semester Assessment Practice Questions for Class VIII Indonesian Language Subjects at SMPN I Bambanglipuro Bantul Using the ITEMAN Program. *Kembara: Scientific Journal of Language, Literature and Teaching*, Vol. 8, No. 1, 160-180. E-ISSN: 2442-9287

Kartiningsih, E.D. (2015). Guide to Preparing Literature Studies. Institute for Research and Community Service, Majapahit Mojokerto Health Polytechnic.

Magdalena, I., Fauziah, S. N., Faziah, S. N., Nupus, F. S. (2021). Analysis of Validity, Reliability, Level of Difficulty, and Differential Power of Final Semester Exam Questions for Theme 7 Class III SDN Karet 1 Sepatan. *Star: Journal of Education and Science*, Vol. 3, No. 2; 198-214.

Mustofa, S. (2017). Analysis of TOEFL Test Items (Test of Arabic as a Foreign Language). *Maulana Malik Ibrahim State Islamic University Malang*.

Permatasari, Arvynda. (2014). Management of Online Evaluation of Student Learning Results. *Journal of Educational Management*. Vol. 24 No. 3.

Qomariyah, Lailatul. (2022). Analysis of the Level of Difficulty and Differentiating Power of TOAFL Questions at Hasyim Asy'ari University Tebuireng Jombang. *Oral Arabiya: Journal of Arabic Language Education*, Vol.6, No.1. E-ISSN 2623-2588. Doi:10.32699/liar.v6il.2549.

Rahman, T., Hamdu, G., Putri, A. R., Merliana, A. (2022). Technical Training on Analysis of Assessment Instruments Using Rasch Modeling for Elementary School Teachers. *Abdidas Journal*, Vol. 3, No. 6, 1029-1037.

Ratni. (2023). Analysis of the Level of Difficulty, Differentiating Power, and Effectiveness of Distracting Indonesian Language Test Questions in Elementary Schools. *Prose: Journal of Primary School Teacher Education Research*, Vol.1, No.1. Two: https://doi.org/10.35326/prosa.v8i4.4234

Saputra, H. D., Purwanto, W., Setiawan, D., Fernandez, D., Putra, R. (2022). Student Learning Outcomes: Test Item Analysis. *Education: Journal of Education*, Vol. 20, No. 1, 15-27.

Sari, V.N.I., Utomo, A.P.Y., Sumarwati. (2022). Quality of Indonesian Language Questions at SMP Muhammadiyah 1 Pontianak: Analysis of Question Items. *Journal of Indonesian Language and Literature Education*, Vol.11, No.2. e-ISSN 2503-3476

Son, A. L. (2019). Instrumentation of Mathematical Problem Solving Ability: Analysis of Reliability, Validity, Level of Difficulty, and Differential Power of Question Items. *Echo of Wiralodra*, Vol. 10, No. 1, 41-52. E-ISSN: 2622-2969

Sudijono, Anas. (2019). Introduction to Educational Evaluation (3 ed.). In *MoDuluS: Civil Science World Communication Media*, Vol. 1, Issue 1

Susanto, H., Rinaldi, A., Novalia. (2015). Validity Reliability Analysis of Level of Difficulty and Differentiating Power on Odd Semester Final Exam Questions for Mathematics Class XII IPS at SMA Negeri 12 Bandar Lampung 2014/2015 Academic Year. *Al-Jabar: Journal of Mathematics Education*, Vol.6, No.2. TWO: http://dx.doi.org/10.24042/ajpm.v6i2.50

Valen, A., & Satria, T. G. (2021). Analysis of the Level of Difficulty of PAS Questions (Final Semester Assessment) of Social Sciences Subjects in Elementary Schools. *Basicedu Journal*, Vol. 5, No. 4, 2199–2208. https://doi.org/10.31004/basicedu.v5i4.1149

Wulan H.J.A., & Aristia, R. (2018). Types of Instruments in Learning Evaluation. *Muhammadiyah University of Sidoarjo*.

Wiyasa, P.I., Laksana, I.K.D., Indrawati, N.L.K.M. (2019). Evaluating Quality of Tearcher-Developed English Test in Vocational High School: Content Validity and Item Analysis. *Education Quarterly Review*, Vol. 2, No. 2. SSRN: https://ssrn.com/abstract=3415235