DIFFICULTY DEGREE OF ENGLISH TEST DESIGNED BY HIGH SCHOOL TEACHER

Sahrun

sahrun2408822@gmail.com

English Education Study Program of Universitas Muhammadiyah Palu

Abstract

This study attends to provide information about the difficulty level of English tests in High Schools in Parigi, Central Sulawesi, designed by the teacher. In carrying out the research, a descriptive-quantitative method was applied. The sample of this study is the multiple-choice test for the third-year students of Senior High School in Parigi Selatan consists of 40 items. Based on the findings, the outcomes of the existing data of the test reported that 14 items were classified as difficult, 26 items were classified as moderate, and no items were classified as easy. Based on the results, the researchers conclude that the difficulty of the English test designed by the English teacher of third-year students at SMK Negeri 1 Parigi Selatan is classified as moderate, with a mean score of 0,31.

Keywords: Teaching strategy, Freedom Writers, Analysis.

BACKGROUND

The result of the teaching and learning process is affected by several factors. One of the factors is students' appraisal of the process of evaluation. This appraisal mainly aims at ascertaining the extent of the instructional goals met.

Evaluation activity at any level of education is conducted in the form of tests such as a semester or National examination at the end of every level of education. Astin (2012) supports this by stating that practically everybody in the academic community is evaluated these days. The activity immediately aims at providing necessary information concerning students' performance on any subject matter. So, carefully collected evaluations can help teachers understand their students learning to decide what instructional objectives are being achieved.

Since evaluation procedures provide more reliable data for making such judgments, they should be well prepared. Concerning this statement, Mohan (2016) states that testing and assessment should be designed, monitored, and evaluated for continuous progress as a feed-forward mechanism to judge and improve the triad. All measurement tools should be used to meet the characteristics of an excellent satisfactory procedure.

As one of the assessment forms, a test is frequently held to evaluate students to determine their achievement in learning. Written assessment tools, objective organized clinical tests, and simulated patient-based examinations are examples of standard evaluation tools (Schuwirth, Lambert & Vleuten, 2018). A test mainly aims at observing the whole instructional results.

Teachers' attention is so considerable in justifying or assigning a grade that most teachers pay little attention to the test they use and often ignore it. Dealing with item analysis, it seems teacher-made tests, in general including tests of English, are not sure of how difficult or easy they are, whether or not the item of the test is capable of separating the better from, the poorer students, and how effective each distractor is attracting the students.

Likewise, in other schools in Indonesia, SMK Negeri 1 Parigi Selatan also provides a test to evaluate students' ability after following a subject for the whole semester. The English teacher gives tests to the students for the same purpose. The difficulty degree of the tests should be examined more regularly. Therefore, the researcher is interested in finding out the test quality.

From the situation explained, the research question formulated to focus this research is: What is the difficulty level of the English test designed by the English teacher for third-year students of SMK Negeri 1 Parigi Selatan? From this statement, the objective of the research is stated: it is to identify the difficulty level of the English test designed by the English teacher in the third year of SMK Negeri 1 Parigi Selatan. This research is limited to analyzing the teacher's multiple-choice English tests, which the English teacher makes.

Basic Terms in Language Testing

Evaluation is something familiar in our society. Most people have heard and seen it. Nevertheless, some need help understanding the test, measurement, and evaluation.

1. Test

   There are many meanings for the test. According to Kubiszyn & Borich (2016), a test is a tool that can contribute to evaluating pupils, curricula, and teaching methods. At the same time, Sheeba (2017) adds a more specific definition of tests, especially in education and psychology, as an effort to measure an individual's knowledge, intelligence, or other characteristics in a systematic way. According to Reynolds et al. (2021), test results from psychological tests are highly significant since they serve as the foundation for interpreting an examinee's performance. Testing offers valuable information or inputs regarding the development and achievement of learners' challenges, learning styles, and anxiety levels, and it becomes an essential component of teaching.

2. Measurement.

   Measurement has a slight difference to test. Adom et al. (2020) place measurement and testing as parts of the evaluation in a conceptual framework of differentiating test, measurement, and evaluation. In most cases, measurement involves assigning quantifiable data using one or more tools, such as a test or rating scale. Cizek et al. (2019) explain that a student's knowledge, skills, capacities, and interests are just a few traits that can be learned about through the science and practice of educational measurement. According to Amos et al. (2021), measurement is the process of quantifying the effectiveness and efficiency of an action. From these explanations, measurement emphasizes quantified treatment success and competence data.

3. Evaluation

   Evaluation is considered more comprehensive than measurement. Hakim & Irhamsyah (2020) state that evaluation is an activity that includes two elements, measuring and assessing. According to Rakhmat & Rasyid (2021), the evaluation process is one of the key elements that teachers need to comprehend. It is a crucial part of the process educational component. Evaluation calls for efforts to ascertain the degree to which defined educational goals are achieved.

By this view, the test constitutes an integral part of the measuring process. That is because it provides information about students' intended learning outcomes. Thus, it is clear that the test refers to a systematic procedure containing a set of tasks or questions used to provide information about student performance being supposed to process.

Teacher Made Test

In schools, as evaluators, teachers are assigned to create a test to evaluate their students' achievement after the teaching and learning process is conducted in a class. This test is called a teacher-made test. Arikunto (2013) states that a teacher arranges a teacher-made test without contributions from another party or an expert. It is usually an achievement test, which is only used in criteria school or class.

Different from the standardized test, the teacher-made test is a test that the teacher, he or herself constructs. It has no characteristics as the standardized test and has yet to be tried. In addition, a teacher-made test may also provide the teacher with information or data about the effectiveness of classroom instruction. The teacher can only decide on students with this information.

Following Arikunto (2013), a teacher-made test is suitable for formative and summative evaluations. According to Zook (2021), formative assessments are examinations and quizzes that gauge how well a student understands a subject during a course. Summative assessments, on the other hand, consist of tests and quizzes that measure how much learning has occurred throughout the course. Based on the explanation, formative tests are commonly used to know students' mastery of the materials taught. In order to get a clearer picture of this student-absorptive capacity, the teacher should own a specific criterion to determine the extent to which students have achieved the instructional objectives. In this case, a criterion reverence test is used. Summative tests, in contrast, are used to determine students' grades at every last period of the instructional process (a quarter monthly), evaluate whether the students are capable of completing the test, promote them to a high degree, and indicate their achievement within a classroom.

Standardized Test

A standardized test is different from Teacher Made Test. Kubiszyn & Borich (2016) explain that standardized tests are often created over protracted periods by test construction experts at significant expenditure to standardized test publishers and state education organizations. This test is administered to students consistently, meaning that all the questions are the same, each student gets the same amount of time, and the scoring is done the same way for everyone.

In terms of students' position in learning, the standardized test intended to compare one student to another both in a classroom or inside and outside the school. In terms of the student's progress in learning, it is intended to recognize the achievement they have during the teaching and learning process is taking place. In this case, the teacher often administers it to the student. Furthermore, the diagnostic is intended to diagnose the student's learning weakness and improve it so that the established instructional objectives can be achieved.

Test Item Analysis

An item analysis is a process to obtain data about test values. According to Ani (2011), item analysis is a systematic procedure by which the teacher can get some information about the quality of the test item. Meanwhile, Madsen (1983:180) stated that more than selecting appropriate language items are needed to ensure a good test. Each question needs to function correctly. Otherwise, it can weaken the exam. Fortunately, there are some relatively simple statistical ways of checking an individual's item. This procedure is called "item analysis ." It is most often used with multiple-choice questions.

An item analysis tells us three things: how difficult each item is, whether or not the question "discriminates" or tells the difference between high and low students, and which dictators are working as they should. An analysis like this is used with any actual exam-for example, review tests and tests are given at the end of a school term or course. To prepare for the item analysis, first score all of the tests. Then arrange them from the one with the highest score to the one with the lowest. Next, divide the papers into three equal groups: those with the highest scores in one stack and the lowest in another. (The classical procedure is to choose the top 27 percent and the bottom 27 percent of the papers to analyze. However, since language classes are usually relatively small, dividing the papers into thirds gives us essentially the same results and allows us to use a few more papers in the analysis). In addition, Madsen (1983:178) stated that besides being on the right level and covering the material discussed in class, good tests are also valid and reliable.

A valid test measures what it claims to be measuring. A reliable test produces the same results consistently on different occasions when the test conditions remain the same. Therefore, item Analysis is related to the several items of statistical analysis in analyzing characteristics and features of a test. They consist of validity, reliability, and level of difficulty.

1. Validity

The validity of a test is the first characteristic of a good test. Suppose it has validity in measuring the students. Sudaryono et al. (2019) explain that validity means the extent of the determination and accuracy of a measuring instrument in performing its measuring function. A good test should possess validity: it should measure what it is intended to measure and nothing else. The test provides information about the general subject matter or source that should be regulated.

2. Reliability

Reliability is the second characteristic of a good test. In this case, reliability means the stability of the test score. A test can only measure something well if it measures consistently. Livingston (2018) states that reliability is the extent to which test scores are not affected by chance factors by the luck of the draw. Suppose a test puts several students in a different order of merit when administered a second time (provided that neither teaching nor learning has taken place in the interval). In that case, the test needs to be more reliable. Reliability is a necessary characteristic of any good test: for it to be valid at all, a test must first be reliable as a measuring instrument.

3. The level of items difficulty

Determine item difficulty on a test related to the number of students who correctly answered the item. A good test has to be a proportional item, which means a test has to have to balance within an easy item, enough, and the difficulty. According to Ani (2011), doing the test item analysis can help teachers determine that the items effectively evaluate students' learning progress.

The excellent item consists of several difficult, fair, and easy items. A very easy test will not stimulate to enrich the students' effort to solve it. While a test that is too difficult will make students hope less, and there is no passion for trying again. In determining item difficulty can do as follows:

Step 1: separate the highest and the lowest 25% of the papers.

Step 2: For each item, subtract the number of “lows” who answered the item correctly from the number of “highs” who answered correctly.

Step 3: Divide the result of step 2 by the number of papers in each group.

In analyzing the item to identify a good, satisfactory, and poor item, the researcher used the formula of the difficulty level. Ani (2011) adds that item analysis helps teachers and test developers evaluate students’ learning competence, which interprets students' progress at the end of the learning process. At this point, the degree of student capacity in answering an item might influence the difficulty level (Hasan and Zainul, 1991/1992). If an item is administered to the two groups of students, the difference in their ability and test results tend to be different. Thus, the item difficulty level is intended not to show which items are poor or good but indicate which items are easy for the students who tried the test or which are difficult for them.

Good items of each test should be attempted to put the proportion in the middle of the range. However, it is impossible that difficulty values always fall in the proportion of 0.5; the possibility is somewhere in the range from zero to one. For simplicity, they are classified in difficult order into a difficult-moderate or easy category. For this view, it is suggested that the teacher is hoped to use the following criteria for an estimate in constructing a test with a multiple-choice item: the difficult items ranging from 0.00 to 0.24, the moderate ones falling somewhere between 0.25 and 0.75: and the easy ones ranging from 0.76 to 1.00 (Hasan and Zainul, 1991/1992). Thus, a good item should range from 0.25 to 0.75.

METHOD OF THE RESEARCH

In terms of method, this is descriptive research. Gay et al. (2012) state that survey research is descriptive research. It means it describes the things on the field the way it is. This research intends to describe the characteristics of items tested in difficulty level.

Two variables involved in this research are independent and dependent. The independent variable revers the English test designed by the English teacher. The dependent variable reverses the difficulty degree of the English test in terms of item difficulty level, item discrimination, and the effectiveness of distractors.

The population of this research was English tests used at SMK Negeri 1 Parigi Selatan in the third year. This research applies the cluster random sampling technique. The research sample was multiple choices given in the academic year 2019/2020. It consists of 40 items.

In carrying out this research, the writer used two instruments:

Two sets of English tests were from the answer sheets of students. The keys of tests and scoring system in multiple-choice prepared by the English teacher.

The desired data was collected using the following procedure:

1) Visiting and asking permission from the English teacher as well as the head Master of the school;
2) Collecting the test result of students;
3) Listing the students who responded with either correct or incorrect item in the table;
4) Applying the procedure of item analysis as follows:
   - Ranking the answer sheet paper from the higher to the lower scores.
   - Computing the difficulty of each item.
   - Tabulate the number of students in the lower and higher group who respond correctly and incorrectly. If the students respond incorrectly, give – (minus); if they respond correctly, give + (plus).
   - Add the number of students in the higher and the lower groups who respond incorrectly in the tabulation.
   - Apply the formula of the item difficulty index.

The collected data was then analyzed to determine the item characteristics in terms of difficulty:

Item Difficulty Index

$$\mathrm{P} = \frac{R}{T}$$

Where:

P = Index of the item difficulty

R = the number of students who got the item right

T = the total number of students who tried the item

Grondlund (1976: 276)

The purpose of data interpretation the collection data is to indicate the item characteristics in terms of difficulty, discrimination, and effectiveness of distractor using the following criterion as shown in the table below:

No.	Value	Criterion
1.	0.00-0.25	difficult
2.	0.26-0.75	moderate
3.	0.76-1.00	easy

Source: Grondlund (1976: 280)

FINDING AND DISCUSSION

This research investigates the English test used at SMK Negeri 1 Parigi Selatan. The test consists of 40 items and involves 42 students. Analyzing students' scores on each test item can be seen in the test result.

Table Classification of test items in difficulty level

CRITERION	CLASSIFICATION
Difficult (0.06 – 0.25)	4, 10, 11, 15, 16, 17, 19, 21, 22, 23, 26, 27,30, 38.
Moderate (0.26– 0.75)	1, 2, 3, 5, 6, 7, 8, 12, 13, 14, 16, 18, 20, 24, 25, 28, 29, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40.
Easy (0,76-1.00)	-

The table shows 14 items classified as difficult, 26 items classified as moderate, and no items classified as easy.

DISCUSSION

This part is in line with the interpretation of the findings derived from the previous quantitative analysis. Based on the findings, the outcomes of the existing data of the test reported that 14 items were classified as difficult, 26 items were classified as moderate, and no items were classified as easy.

This fact provides us with a point about the current condition of the English test used for the third-year students at SMK Negeri 1 Parigi Selatan. Arikunto (2013:222) stated that a good test is a test that is not too easy or too difficult. A good test item should have a difficulty level, including easy, moderate, and difficult. A practical and good test should have items that belong to a moderate level. The too-easy or difficult item weakens the test quality, and valid data about students' achievement will not be acquired.

CONCLUSION AND SUGGESTION

Conclusion

A test becomes an evaluation activity conducted at the end of every level of education in the form of a semester or National examination. The difficulty level of the test affects the evaluation result. The researcher analyzed the difficulty level of the English test designed by the teacher. As a result, it is found that the mean score of the test item difficulty index is 12,69/40 = 0,31. This means the difficulty level of the test used in SMK Negeri 1 Parigi Selatan is categorized as moderate.

Suggestion

After conducting this research, there are several suggestions raised for the improvement of English teachers' ability in designing tests and also in the difficulty level of tests given to students. They are:

- So that teacher-made tests fulfill the criteria as a good level test, classroom teachers in general including teachers of

English should know what should be tested. Since the blueprint of tests teachers prepare or develop tells the instructional objectives or topics to be covered, the cognitive levels to be involved, and the number of items to be provided.

- After having the students' test results, the teachers must analyze their test items in terms of difficulty level.
- The school headmaster actively guides and upgrades the teacher concerning how to develop a test correctly, especially a good test in difficulty level.

REFERENCES

Adom, D., Adu-Mensah, J., & Dake, D. (2020). Test, measurement, and evaluation: Understanding and use of the concepts in education. *International Journal of Evaluation and Research in Education (IJERE)*, *9*(1), 109–119. https://doi.org/10.11591/ijere.v9i1.20457

Amos, D., Au-Yong, C. P., & Musa, Z. N. (2021). *Measurement of Facilities Management Performance in Ghana's Public Hospitals*. Springer.

Ani, L. A. (2011). *An Item Analysis on The Difficulty Level of an English Summative Test for Second Grade of SMP Muhammadiyah 29 Cinangka-Sawangan Depok* [Syarif Hidayatullah State Islamic University Jakarta]. https://repository.uinjkt.ac.id/dspace/bitstream/123456789/4878/1/100839-LIA ANDRI ANI-FITK.PDF

Arikunto, S. (2013). Prosedur Penelitian: Suatu Pendekatan Praktik. Jakarta: Rineka Cipta.

Astin, A. W. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Rowman & Littlefield Publisher.

Cizek, G., Agger, C., & Lim, S. (2019). *Measurement in Education in the United States*. Oxford Bibiliographies. https://doi.org/10.1093/OBO/9780199756810-0060

Grounlund . N.E. 1976/1968. *Measurement and Evaluation in Teaching. New York. Mac-Milan.*

Hakim, L., & Irhamsyah, I. (2020). the Analysis of the Teacher-Made Test for Senior High School At State Senior High School 1 Kutacane, Aceh Tenggara. *JURNAL ILMIAH DIDAKTIKA: Media Ilmiah Pendidikan Dan Pengajaran*, *21*(1), 10. https://doi.org/10.22373/jid.v21i1.4120

Hasan, Hamid S., dan Zainul, Asmawi (1991). Evaluasi Hasil Belajar. Jakarta:
Depdikbud.

Kubiszyn, T., & Borich, G. D. (2016). *Educational testing and measurement* (11th ed.). John Wiley & Sons.

Livingston, S. A. (2018). Test Reliability - Basic Concepts. *Research Memorandum ETS RM-18-01, January*, pp. 1–38. https://www.ets.org/Media/Research/pdf/RM-18-01.pdf

Madsen, H.S. 1983: Techniques in testing. New York and Oxford: Oxford University Press. viii + 212 pp. ISBN 0-19-434132-1.

Mohan, R. (2016). *Measurement, Evaluation, and Assessment in Education*. PHI Learning Private Limited.

Rakhmat, A. T., & Rasyid, A. F. (2021). Word Exploration and the Meaning of Evaluation n the Holy Quran (Thematic Interpretation of Education Evaluation). *Religio Education*, *1*(1), 15–24. https://doi.org/https://doi.org/10.17509/re.v1i1

Schuwirth, Lambert, W. T., & Vleuten, C. P. M. van der. (2018). How to Design a Useful Test: The Principles of Assessment. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding Medical Education:*

Evidence, Theory, and Practice. https://doi.org/https://doi.org/10.1002/9781119373780.ch20

Sheeba, S. (2017). Importance of Testing in Teaching and Learning. *International Journal of Society and Humanities*, *2*(1), 1–9. https://www.researchgate.net/publication/328355159_Importance_of_Testing_in_Teaching_and_Learning

Sudaryono, Rahardja, U., Aini, Q., Isma Graha, Y., & Lutfiani, N. (2019). Validity of Test Instruments. *Journal of Physics: Conference Series*, *1364*(1), 12050. https://doi.org/10.1088/1742-6596/1364/1/012050

Zook, C. (2021). Formative vs. Summative Assessments: What's the Difference? *Applied Educational System*, 1–8. https://www.aeseducation.com/blog/formative-vs.-summative-assessments-what-do-they-mean%0Ahttps://www.aeseducation.com/blog/formative-vs.-summative-assessments-what-do-they-mean#:~:text=In a nutshell%2C formative assessments have learned throughout a course